Tracking domain knowledge based on segmented textual sources

DISSERTATION

Zur Erlangung des akademischen Grades
doctor rerum politicarum
(Doktor der Wirtschaftswissenschaft)

eingereicht an der

Wirtschaftswissenschaftlichen Fakultät
der Humboldt-Universität zu Berlin

von

Dipl.-Kfm. Tobias Kalledat
(geb. am 29.02.1972 in Berlin)

Präsident der Humboldt-Universität zu Berlin:
Prof. Dr. Dr. h.c. Christoph Markschies

Dekan der Wirtschaftswissenschaftlichen Fakultät:
Prof. Oliver Günther, Ph.D.

Gutachter: 1. PD Dr. Bernd Viehweger
2. Prof. Dr. Myra Spiliopoulou
3. Prof. Dr. Anke Lüdeling

Tag des Kolloquiums: 10.02.2009

Zusammenfassung

Text Data Mining (TDM) entwickelte sich innerhalb der vergangenen Jahre zu einem etablierten Forschungsfeld. Es bedient sich eines Kanons von Methoden aus mehreren Disziplinen, mit dem Ziel neues Wissen durch die Anwendung von Data Mining Prozessschritten aus Textkorpora verschiedener Art zu generieren. Dieser Prozess besteht im Wesentlichen aus den Schritten *Datenauswahl, Datenvorverarbeitung, Transformation, Data Mining und Auswertung/Interpretation*. Während bei angewandten Data Mining Vorhaben der höchste zeitliche Aufwand in die ersten zwei vorverarbeitenden Phasen investiert wird, besteht ein Mangel an Forschung über den Einfluss unterschiedlicher Qualitätsniveaus der Vorverarbeitung auf die Qualität des generierten Wissens sowie quantitative Indikatoren für “gut vorverarbeitete” Korpora. Die hier vorliegende Forschungsarbeit hat zum Ziel, Erkenntnisse über den Einfluss der Vorverarbeitung auf die Ergebnisse der Wissensgenerierung zu gewinnen und konkrete Handlungsempfehlungen für die geeignete Vorverarbeitung von Textkorpora in TDM Vorhaben zu geben.

Der Fokus liegt dabei auf der Extraktion und der Verfolgung von Konzepten innerhalb bestimmter Wissensdomänen mit Hilfe eines methodischen Ansatzes, der auf der waagerechten und senkrechten Segmentierung von Korpora basiert. Ergebnis sind zeitlich segmentierte Teilkorpora, welche die Persistenzeigenschaft der enthaltenen Terme widerspiegeln. Innerhalb jedes zeitlich segmentierten Teilkorpus können jeweils Cluster von Termen gebildet werden, wobei eines diejenigen Terme enthält, die bezogen auf das Gesamtkorpus nicht persistent sind und das andere Cluster diejenigen, die in allen zeitlichen Segmenten vorkommen.

Auf Grundlage einfacher Häufigkeitsmaße kann gezeigt werden, dass allein die statistische Qualität eines einzelnen Korpus es erlaubt, die Vorverarbeitungsqualität zu messen. Vergleichskorpora sind nicht notwendig. Die Zeitreihen der Häufigkeitsmaße zeigen signifikante negative Korrelationen zwischen dem Cluster von Termen, die permanent auftreten, und demjenigen das die Terme enthält, die nicht persistent in allen zeitlichen Segmenten des Korpus vorkommen. Dies trifft ausschließlich auf das optimal vorverarbeitete

Korpus zu und findet sich nicht in den anderen Test Sets, deren Vorverarbeitungsqualität gering war. Werden die häufigsten Terme unter Verwendung domänenspezifischer Taxonomien zu Konzepten gruppiert, zeigt sich eine signifikante negative Korrelation zwischen der Anzahl unterschiedlicher Terme pro Zeitsegment und den einer Taxonomie zugeordneten Termen. Dies trifft wiederum nur für das Korpus mit hoher Vorverarbeitungsqualität zu. Eine semantische Analyse auf einem mit Hilfe einer Schwellenwert basierenden TDM Methode aufbereiteten Datenbestand ergab signifikant unterschiedliche Resultate an generiertem Wissen, abhängig von der Qualität der Datenvorverarbeitung.

Mit den in dieser Forschungsarbeit vorgestellten Methoden und Maßzahlen ist sowohl die Qualität der verwendeten Quellkorpora, als auch die Qualität der angewandten Taxonomien messbar. Basierend auf diesen Erkenntnissen werden Indikatoren für die Messung und Bewertung von Korpora und Taxonomien entwickelt sowie Empfehlungen für eine dem Ziel des nachfolgenden Analyseprozesses adäquate Vorverarbeitung gegeben.

Schlagwörter:

Text Data Mining, Korpuskennzahlen, Korpuslinguistik, Computerlinguistik, Datenvorverarbeitung, Vorverarbeitungsqualität, Wissensextraktion

Abstract

During recent years text data mining (TDM) has become a well-established research field. It uses a canon of methods from several disciplines with the aim of generating new knowledge by the application of a “standard” data-mining process out of textual data that is available as different kinds of text corpora. This process consists of the steps of *data selectio*n, *data pre-processing*, *transformation*, *data mining* and *evaluation/interpretation*. Whereas the highest effort needs to be applied to the first two preparing phases, a lack in research is to be found in the analysis of the influence of different quality levels of pre-processing on extracted knowledge and the creation of measures for “well pre-processed” corpora. The research work available here has the goal of analysing the influence of pre-processing on the results of the generation of knowledge and of giving concrete recommendations for action for suitable pre-processing of text corpora in TDM.

The research introduced here focuses on the extraction and tracking of concepts within certain knowledge domains using an approach of horizontally (timeline) and vertically (persistence of terms) segmenting of corpora. The result is a set of segmented corpora according to the timeline. Within each timeline segment clusters of concepts can be built according to their persistence quality in relation to each single time-based corpus segment and to the whole corpus.

Based on a simple frequency measure it can be shown that only the statistical quality of a single corpus allows measuring the pre-processing quality. It is not necessary to use comparison corpora. The time series of the frequency measure have significant negative correlations between the two clusters of concepts that occur permanently and others that vary within an optimal pre-processed corpus. This was found to be the opposite in every other test set that was pre-processed with lower quality. The most frequent terms were grouped into concepts by the use of domain-specific taxonomies. A significant negative correlation was found between the time series of different terms per yearly corpus segments and the terms assigned to taxonomy for corpora with high quality level of pre-processing. A semantic analysis based

on a simple TDM method with significant frequency threshold measures resulted in significant different knowledge extracted from corpora with different qualities of pre-processing. With measures introduced in this research it is possible to measure the quality of applied taxonomy. Rules for the measuring of corpus as well as taxonomy quality were derived from these results and advice suggested for the appropriate level of pre-processing.

Keywords:

Text Data Mining, Corpus Measures, Corpus Linguistics, Computational Linguistics, Data Pre-processing, Pre-processing Quality, Knowledge Extraction

Table of contents

Zusammenfassung 2

Abstract 4

Widmung 12

Abkürzungsverzeichnis 13

Preface 15

1 Motivation 20

1.1 Aim of research 21

1.2 Document structure 24

2 Introduction 26

2.1 What is a text? 31

2.2 What are “Trends” and “Hypes”? 34

2.3 Challenges with progress extraction from text approaches 36

2.4 Implications on current work 39

3 Domain progress extraction 40

3.1 Text source selection 42

3.1.1 Methods to exploit domain knowledge in the mining process 43

3.1.2 Excurse: Underpinnings of evolution in business informatics magazine titles 47

3.1.3 Text Source 1: WWW Archive of German “Computerwoche” 48

3.1.3.1 Semantic benchmark for text source 1 50

3.1.4 Text Source 2: Printed Allianz Management Reports 51

3.1.4.1 Semantic benchmark for text source 2 52

3.1.5 Remarks to the semantic benchmark 53

3.2 Introduction of relevant aspects and methods for the TMF process 54
3.2.1 A cost function for domain knowledge extraction 54
3.2.2 Methods for data-quality evaluation 55
3.2.3 Text data mining 56
3.2.3.1 Clustering and naming 60
3.2.3.2 Progress extraction and topic evolution 62
3.2.3.3 Topic detection and tracking 63
3.2.3.4 Literature mining 65
3.2.3.5 Complexity reduction 67
3.2.3.6 Semantic evolution 67
3.2.3.7 Human-driven methods 69
3.2.4 Computational linguistics 70
3.2.4.1 Text (data) mining and computational linguistics 72
3.2.5 Knowledge representation 75
3.3 Pre-processing 78
3.3.1 The method used here 80
3.3.2 Pre-processing of CW 80
3.3.3 Pre-processing of AI1k 83
3.4 Conversion into a standard format 83
3.5 Pre-Filtering and corpus measure pattern recognition 83
3.5.1 Task-specific segmentation of text collections 84
3.5.2 Corpus measure based domain progress extraction paradigm 88
3.5.3 Corpus measure selection 89

3.5.4 Discussion: Implications of TRQ value as threshold 94

3.6 Data processing and text (data) mining 96

3.6.1 Taxonomy construction 97

3.6.2 Decomposition of constant domain-related and language-related terms 98

3.6.2.1 Discussion: Qualities of volatile domain-related terms 99

3.6.3 TDM on segmented corpora based on TRQ threshold 100

3.7 Domain knowledge interaction 102

3.7.1 Visualization approaches 104

4 Empirical results and evaluation 106

4.1 Observed determining factors on knowledge extraction 107

4.2 Data models 109

4.2.1 Dimensions 110

4.2.2 Measures 116

4.3 Evaluating the impact of intensity of pre-processing 117

4.3.1 Corpus type n (high pre-processing intensity) 122

4.3.1.1 Statistical analysis of type n corpora 122

4.3.1.1.1 Descriptive statistics of CW corpus test set CW_{5k} 122

4.3.1.1.2 Descriptive statistics of CW corpus test set CW_{1k} 123

4.3.1.1.3 Descriptive statistics of Allianz corpus test sets $Al1k_{S1}$ and $Al1k_{S2}$ 124

4.3.1.1.4 Excurse: Predictability of the TRQ Plot 126

4.3.1.2 Statistical analysis of type n corpora summary 128

4.3.1.3 Distribution Analysis of applied Taxonomies on type n corpora 129

4.3.1.3.1 Distribution analysis of CW corpus test set CW_{5k} 130

4.3.1.3.2 Distribution analysis of CW corpus test set CW_{1k} 134

4.3.1.3.3 Distribution analysis of Allianz corpus test sets $Al1k_{S1}$ and $Al1k_{S2}$ 138

4.3.1.4 Distribution analysis of applied taxonomies on type n corpora summary 143

4.3.1.5 Semantic analysis of type n corpora 143

4.3.1.5.1 Semantic analysis of CW corpus test set CW_{5k} 143

4.3.1.5.2 Semantic analysis of CW corpus test set CW_{1k} 148

4.3.1.5.3 Semantic analysis of Allianz corpus test sets $Al1k_{S1}$ and $Al1k_{S2}$ 157

4.3.1.6 Semantic analysis of type n corpora summary 160

4.3.2 Evaluation of Corpus type b (low pre-processing intensity) 160

4.3.2.1 Statistical analysis of type b corpora 160

4.3.2.1.1 Descriptive statistics of CW corpus test set CW_{5kb} 160

4.3.2.1.2 Descriptive statistics of CW corpus test set CW_{5kbu} 161

4.3.2.1.3 Descriptive statistics of CW corpus test set CW_{5kbun} and CW_{5kbun2} 162

4.3.2.1.4 Descriptive statistics of CW corpus test set CW_{1kb} 164

4.3.2.1.5 Descriptive statistics of CW corpus test set CW_{1kbu} 165

4.3.2.2 Statistical analysis of type b corpora summary 166

4.3.2.3 Distribution analysis of applied taxonomies on type b corpora 166

4.3.2.3.1 Distribution Analysis of CW corpus test set CW_{5kb} 167

4.3.2.3.2 Distribution Analysis of CW corpus test set CW_{5kbu} 170

4.3.2.3.3 Distribution Analysis of CW corpus test sets CW_{5kbun} and CW_{5kbun2} 173

4.3.2.4 Distribution analysis on type b corpora summary 177

4.3.2.5 Semantic analysis of type b corpora 178

4.3.2.5.1 Semantic analysis of CW corpus test set CW_{5kb} 178

4.3.2.5.2 Semantic analysis of CW corpus test set CW_{5kbu} 186

4.3.2.5.3 Semantic analysis of CW corpus test set CW_{5kbun} and CW_{5kbun2} 194

4.3.2.6 Semantic analysis of type b corpora summary 217

4.4 Evaluating the impact of language of origin 218

4.4.1 Language fingerprint on corpus level 218

4.4.2 Language fingerprint on corpus-level summary 222

4.4.3 Language fingerprint on concept level 223

4.4.4 Language fingerprint on concept-level summary 230

4.4.5 Analysis of statistical indicators for German corpus subsets 231

4.4.6 Analysis of statistical indicators for German corpus subsets summary 240

4.4.7 Analysis of statistical indicators for Al1k English corpus subsets 241

4.4.8 Analysis of statistical indicators for English corpus subsets summary 249

4.5 Evaluating the impact of corpus length 250

4.5.1 Effects on statistical qualities and their measures 250

4.5.2 Effects on quality of extracted knowledge 251

4.5.3 The “minimal” corpus size for the TRQ measure threshold approach 251

4.6 Evaluating the impact of knowledge domain and document source 252

5 Domain knowledge interaction 253

5.1 Navigating the constant concepts of CW_{5k} 253

5.2 Navigating the volatile concepts of CW_{5k} 254

6 Conclusion and perspective 256

Sources 261

Appendix 280

Acknowledgements 348

Eidestattliche Erklärung 349

Widmung

für Ailan und Andor

Abkürzungsverzeichnis

AI	Artificial Intelligence
ASCII	American Standard Code for Information Interchange
BI	Business Intelligence
CL	Corpus Linguistics
CW	Computerwoche
DM	Data Mining
DMP	Data-Mining Process
DQ	Data Quality
DQM	Data Quality Management
DTD	Document Type Definition
GML	Generalized Markup Language
HGB	Handelsgesetzbuch
HTML	Hypertext Markup Language
IT	Information Technology
KD	Knowledge Discovery
KDD	Knowledge Discovery in Databases
KKMDB	Karlsruher Kapitalmarktdatenbank
ML	Machine Learning
MWF	Mean Word Frequency
OWL	Ontology Web Language
RDF	Resource Description Framework
SMS	Short Message Service
SGML	Standard Generalized Markup Language

SOM	Self-organizing maps
TDM	Text Data Mining
TDT	Topic Detection and Tracking
TEI	Text Encoding Initiative
TMF	Trend-Mining Framework
TQC	Total Quality Control
TQM	Total Quality Management
TRQ	Term-Repetition Quota
TTR	Type (to) Token Ratio
WWW	World Wide Web
XML	Extended Markup Language

Preface

The 21st century is proposed to be the *Century of Knowledge* by most of the knowledge-oriented researchers and practitioners around the world (e.g., Goverdhan Mehta[1]). Tim Berners-Lee, as one of the leading creators of the current World Wide Web (WWW), invented a vision of a semantic extension of the WWW that re-uses and combines knowledge from all available sources to build a universal Semantic Web[2]. Business-focused researchers also realize that a fundamental change in the focus of business progress is underway. Nefiodow ([Nefi96], pp. 126]) declared that the 21st century belongs to the 5th Kondratieff cycle of social development. The fifth Kondratieff is, in his interpretation, the first long cycle which came about through the utilization of mineral resources, no longer primarily of processes of substance transformation and energies but of the intellectual utilization of information. Mankind experiences information-driven structural change instead of the energy-driven change of the previous cycle. Information forms the raw material for the development and entrepreneurial use of the factor of production knowledge. An increase of the knowledge-intensive lines of business and tasks has been witnessed in the First and Second World since the middle of the 20th century. More and more enterprises have come into being that no longer produce and trade real goods, rather virtual goods. Typical representatives are software companies, service or consulting firms as well as banks and insurance companies. This has given special importance to the management of that knowledge within enterprises that act in the global market. The implications of information and knowledge on internal structures and processes are fundamental. E.g., [Biss99], p. 376 analysed the implications and perspectives of these developments from a managing perspective. With the term “knowledge” several aspects have to be considered, starting with questions regarding protection of intellectual property by patents and law, free access to sources and ending with methodological tasks, e.g., the creation of description logics for representation purposes. The need for internalis-

[1] President, International Council for Science (ICSU), cited from his session “Science and Technology Capacity and the Knowledge Society” at World Science Forum Budapest Nov 10, 2005

ing new knowledge in a short time is crucial for knowledge workers[3] (see [Druc70], p. 270f). If one is trying to follow up a special knowledge domain, he or she may appreciate not having to read millions of documents, but getting supported in a way that presents the relevant concepts and domain progress in an intuitive manner. The law of diminishing returns is also valid for knowledge: A certain increase of information does not automatically lead to the same ratio of more knowledge.

The internalization of new domain knowledge can be done in different ways: Trying to read a large amount of related texts and talking with time witnesses and specialists in the current domain, but what if there is no one available to answer questions? Collecting empirical data in the form of texts may be appropriate. The greatest difficulties in the collection of empirical data about most knowledge domains can arise from the non-availability of suitable knowledge carriers for an investigation with a long view horizon (e.g., more than 50 years for the IT domain or thousands of years in archaeology). A further problem is that such information is regarded by some enterprises (from the demand side) as strategic and not shared with external researchers. A large range of popular written literary works exists about different personalities and enterprises (success stories, memoirs of great leaders), that have considerably influenced the history of business informatics. However, they are all too often no more than a subjective representation of reality. As supplementary sources of information for primary orientation or as a documentation of the spirit of the time, such documentary issues may be helpful, but they lack neutrality of reporting. Even if used for analysis, such sources must be translated into a computable format to be a basis for quantitative investigations.

---

[2] Theoretical foundations can be found in [Anto04a], pp. 3.

[3] Here the term *knowledge worker* is used, which Peter Drucker described as follows: "The knowledge worker's demands are much greater than those of the manual worker, and are indeed quite different. [...] Knowledge workers also require that the demands be made on them by knowledge rather than by people. [...] Knowledge, therefore, has to be organized as a team in which the task decides who is in charge, when, for what, and for how long."

The example of the knowledge domain “IT” can be seen as a market paradigm in which two parties are acting: Buyers of IT goods that produce a demand, on the one hand, and Sellers of IT goods, which compete against others regarding the best offer for the Buyers of their goods. IT goods in this context may be hardware, software and support (e.g., consulting) of different kinds. For the aim of tracking developments in the IT Domain it would be opportune to ask the buyer or seller side about their historical IT knowledge. However, several problems arise: There are not enough time witnesses available for interviews and the reflection of the historical reality may not be free of subjective information distortions. While acquiring this information it can be learned that the buyer side has no focus on documenting all their IT history, while focusing on their main (non-IT) business. Therefore, the seller side is potentially more capable than the buyer side of giving detailed information of their historical business, in addition to technological development and the associated product portfolio. Concerning the quality of the information placed, it has to be noted that this is subject to a seller-individual preselection and is not given a warranty for completeness, e.g.: In the case of product developments, which were not successfully placed on the market, a type of behaviour that is known as information hiding can occur. This can be explained by the interest of the seller enterprises to present the best possible image to the public.

The problems are much greater if a knowledge domain that exists for hundreds or thousands of years is to be tracked, e.g., law or astronomy. There are no time witnesses available and the focus of research has to turn to available explicit (written) documentation. The advantage of an evaluation of historical text documents in contrast with interviewing time witnesses exists in the reflection of the historical reality free[4] of subjective information distortions. On this assumption, the provision of information problems is reduced to the procurement of suitable text documents. Expenditures for the determination of time witnesses and for the execution and evaluation of interviews can thus

[4] The reflection within a certain text may be author-specific and therefore not free of individual preferences, but if a large document basis is used, the individual biasing of information regarding certain topics, events or persons is not expected to be relevant. That, of course, depends on the variety and number of documents.

be minimized. The analysis of electronic sources also grants the opportunity of transforming semantic structures of knowledge domains in a logical representation that more easily allows for the documenting, sharing and interacting with that knowledge. The current status is characterized by two facts:

a) The production of textual documents increases over time due to the distribution of Information Sources and their shared use, e.g., over the Internet. The availability of internet pages indexed by the search engine a) Google increased from about 1 billion pages in 2000 to more than 8 billion early in 2005 (see Fig. 1). The availability of large sources of potentially interesting knowledge in companies also becomes ubiquitous and the number of produced textual documents rises dramatically.

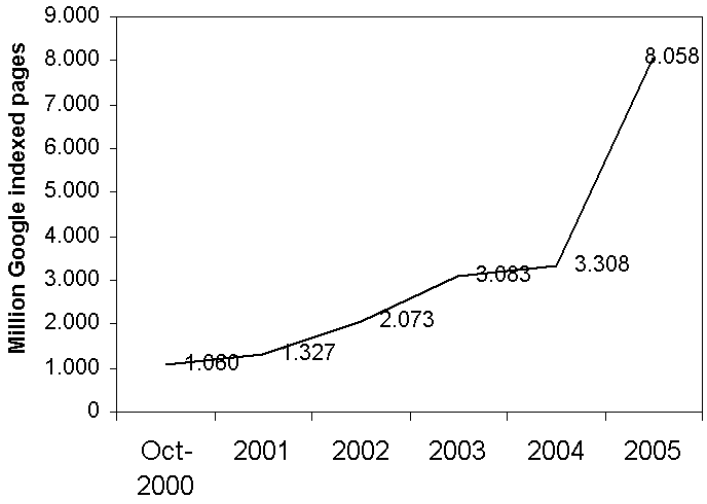


Fig. 1: Google indexed pages (Source: [Capt05])

b) The usage of actual knowledge becomes more and more a critical factor for competition of market participants. To adjust their product portfolio quickly it is necessary to adapt new developments in a short time period, because of consumers asking for product life cycles which are getting shorter and services becoming more adapted to their individual needs. For employees and the management staff of firms, that are encouraged to support the underlying processes, lifelong study becomes more and more important and the flexibility of adapting new domain knowledge in a short time turns out to be one of the most important tasks.

During the last few decades of the 20th century the use of implicit knowledge hidden in the huge amount of unstructured data has become a real option

because of the rapid development of powerful hardware that can handle such large data sources. The methods that were developed under the terms “Knowledge Discovery” (KD) and "Data Mining" (DM) since the early 1990s [Fayy96] are crucial. The Data-mining process permits the discovery of formerly unknown, potentially useful knowledge patterns from various kinds of input data using methods of statistics, linguistics and database research [Düsi99], pp. 345, [Biss99], pp. 375.

For the domain knowledge researcher the following problems of analyzing large amounts of textual data result:

- The pure amounts of unstructured historical and present data represent a substantial entrance barrier.
- Unstructured data cannot be easily automatically processed.
- The found knowledge must be made “visible” for a learning process.

A substantial realization gain is to be expected if methods are found to open and evaluate the mentioned, unstructured sources of information.

It is economically viable to support individuals engaged in knowledge discovery. Potentially expensive manual work can be substituted by automatically working business informatics driven solutions. The budget restriction must always be considered when implementing applications in real-life scenarios. Advice from theoretic research may help here.

1 Motivation

In a world of increasingly expanding information resources, there is a demand for enabling a large amount of knowledge workers in companies and organizations to be able to acquire the knowledge they need out of the ocean of unstructured texts, e.g., from large text archives[5]. For a domain knowledge researcher it is important to know the answers to the following questions: How do the semantics of terms change over time? Which topics are increasing, decreasing? What is the semantic basis of the domain? Rules for significant decisions are needed to distinguish between these clusters. An important challenge for research is to define methods which can extract significant patterns and track time-dependent changes. There is a need for methods to carry out the tasks mentioned above which can properly support the whole knowledge acquisition process. Alternatively applied methods may need different levels of effort to be invested. From real-life experience a positive correlation between the effort invested in processing the source data and the quality of extracted knowledge is to be expected. Limited economical sources force the task of proposing “optimal” intensities of processing.

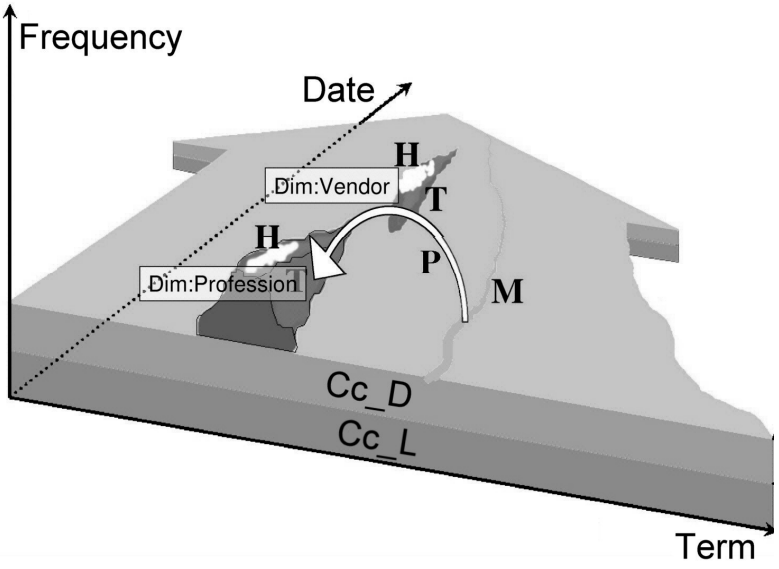


Fig. 2: "Trend Landscape” metaphor

[5] An example of applications in the media industry is introduced by Peters [Pete05b].

Thinking about the ways knowledge domains may be represented led to a blueprint that reminded me of a 3D graphic constructed with the perspectives “Time”, “Term” (or aggregated concepts) and “Frequency” (measured occurrence). This was all composed into the “Trend Landscape” metaphor that is schematically shown in Fig. 2.

Without defining all components shown in the graphics above in detail, the idea behind a “Trend Landscape” for knowledge representation is briefly introduced here. Start by measuring a text collection with quantitative indicators, e.g., frequency of terms or other simple measures, and then project them into a time-dependent representation on an aggregated concept basis, e.g., domain-specific topics. This is what the letters “M” and “P” should illustrate: A time series of measure values – a “flow” in Fig. 2 – which is projected into simple aggregated concepts like “Profession” or “Vendor” that rise as “mountains” over the “ground” of terms. The process of projection may be supported by statistical methods, which allow significant assignments of relevant concepts and filtering of irrelevant concepts. It would then be helpful to differ between concepts that occur only a short time period (so-called hypes, abbr.: as “H”) and longer present trends (abbr. as “T”). Another perspective on domain knowledge may be the differentiation between the constant (persistent) terms within the ground “C_C” that belong to the language “L” and terms that belong to the domain “D” itself. And finally it would be helpful for a discovery of domains if the “Trend Landscape” allowed an interactive navigation within the structure of concepts inclusive of detailing from the concept level down to the term level.

1.1 Aim of research

The aim of the current work here is to define methods for text corpus processing that allow one:

- To recognize the evolution of language and contents (progress path) within a knowledge domain

- To extract characteristic concepts and cluster them into a group of those typical for the domain and others that are only characterized over a short time period as a kind of “fashion”
- To group the concepts in a domain-specific structure

Object of research: Methods for processing textual corpora where these collections are specifically domain-related collections of textual documents of no special origin. In order to find a technical solution that enables the “Trend Landscape” metaphor introduced in the previous chapter, the following subtasks must be processed:

- Decomposition of source text collections into segments of terms belonging to concepts/topics, language or knowledge domain whether they are persistent or not.
- Definition of methods for the detection and extraction of developments within these segments of knowledge-domain-related text segments.
- Development of measures for corpus quality that help to analyse the influence of several factors on quality of extracted knowledge, e.g., different pre-processing strategies and source of input data.
- Visualization of the found results and making them accessible for knowledge domain workers.

A negative definition of the aim of the research may help to avoid misunderstandings. Therefore, the “non-aims” are stated here:

- Building a complete integrated software system
- Re-inventing methods for information retrieval[6]

To attain the research aims, an interdisciplinary approach is proposed. This requires using methods from corpus linguistics (CL), text (data) mining (TDM) and expert domain knowledge for evaluation of results derived from the methods applied.

[6] For an introduction to information retrieval see [Corn04], pp. 162 and [Mili05].

Statistically orientated methods from CL, which only consider single terms, are not appropriate. An interdisciplinary combination of approaches from the corpus linguistics and the growing TDM research field are appropriate for an application on time-segmented corpora[7], which represent the published (made explicit) knowledge of a domain. The main focus here is how corpus measures can support knowledge extraction about domain progress out of textual sources and what influence different pre-processing qualities of the used texts have.

It was found that a time-based *horizontal* segmentation of text collections is appropriate to act as a basis for the extraction of time-dependent domain progresses or certain concepts. First, to avoid distortions by the corpus length dependency of the most corpus measures a normalization of each corpus segment to a common corpus segment length throughout all segments is necessary. By identifying the constant elements that occur in all corpus segments a *vertical* segmentation of the elements can be done. In this step the terms occurring within the text collection are separated in volatile and constant terms. The calculation of simple term repetition quotas for each corpus segment allows finding a basis for statistical time series analysis. Depending on the quality of pre-processing, the source data can be identified to be of good quality or not for following data analysis procedures. It was found that the separated volatile and constant segments of a text collection built good clusters based on a correlation analysis based on the corpus linguistic measures and applied statistical test. The proposed method indicates if a given text collection is well pre-processed or not without comparing reference values.

By definition of thresholds based on these corpus segment measure values, terms that rule for a certain period can be significantly identified. By adding an external domain-related taxonomy or ontology, a structuring according to concepts or topics is possible. Based on this processing the text collection can be made available for an Online Analytical Processing that enables a

[7] 'Time segmented corpora' (in this context) means any press publication, e.g., scientific magazines, which appear regularly (weekly or daily).

knowledge worker to interact with the found domain-related concepts. Several applications in scholarly, scientific and business areas can benefit from the empirical results in measuring and comparing several test sets of mostly very large text collections in the fields of pre-processing text sources for tracking domain knowledge while also making this process more efficient. In such scenarios the economical aspect of finding the optimal level of pre-processing is dominant.

1.2 Document structure

This dissertation is didactically orientated towards an intuitively understandable process of knowledge extraction from textual data. The succession of semantic steps is represented in Table 1.

Table 1: Internal constitution of this text

Contents	Chapter
Introduction of basic terms and methods	2
Introduction of the Trend Mining Framework and related research	3
Evaluation	4
Visualisation and navigation through knowledge domains	5
Further research directions	6

This structure is orientated towards the data-mining steps, introduced in [Fayy96], with adaptations that were useful for the current dissertation. Due to the cross-discipline character of this research only a rough overview of related research is given in the introductory chapter. Detailed references are given in the next chapters introducing the different aspects of the current approach.

Some information regarding the organization of this text is given here. Abbreviations are first used, after introducing them in brackets immediately after the term that is to be abbreviated from this point. Signs in cursive brackets, as in the following example, signal all definitions:

[a] This is a definition.

Formulas are numbered in cursive square brackets:

[1] This is a formula.

Later on, the “I” perspective is used when I want to state that my opinion and experiences are meant. I prefer to do so because I have to take the responsibility for theses that I declare. I do this to ensure that everybody can differ between my original theses and those of other researchers.

2 Introduction

In this chapter the scientific background and state of the art regarding the research aim of this dissertation will be introduced. A more in-depth discussion of certain methods follows in the chapters of this text where the parts of the proposed approach are explained in detail.

The historical documentation of domain knowledge is usually done in the form of unstructured text, picture, audio and video documents that are produced over longer time periods. In this context, structured (e.g., relational data) and semi-structured (e.g., HTML pages) and unstructured documents (e.g., texts) become distinguished regarding the degree of internal structure. The usage of structured languages such as XML for tagging of texts is only at the beginning of its development path and is therefore not to be found for historical documents. Thus in the relevant literature it is assumed that up to 80-90% of electronically stored knowledge is hidden in such unstructured sources [Tan99], [Dörr00], pp. 465.

This development, on the one hand, and market demand on the other involves different challenges for educational and other sectors which make use of such approaches, as well as for developers of information systems that support these processes of knowledge discovery (KD). In the early 1960s terms like “data fishing” were used in the context of criticising badly structured data analysis. Gregory Piatetsky-Shapiro originally used the term “Knowledge Discovery in Databases” in his first workshop in 1989, and this term became more popular in the Artificial Intelligence (AI) and Machine Learning (ML) community.

Since the 1990s (under the term ‘Data Mining’) methods were developed which make it possible to recognize unknown structures in data and derive from it action-relevant and economically useful knowledge, refer to [Codd93]. These methods are based on classical statistic procedures as well as methods of adjacent research fields and were adapted for the employment of appropriate data. KDD and Data Mining make use of methods from several research disciplines and also build a bridge between these disciplines. Data

Mining is not a separate research field, but an intersection of multiple disciplines (see Fig. 3).

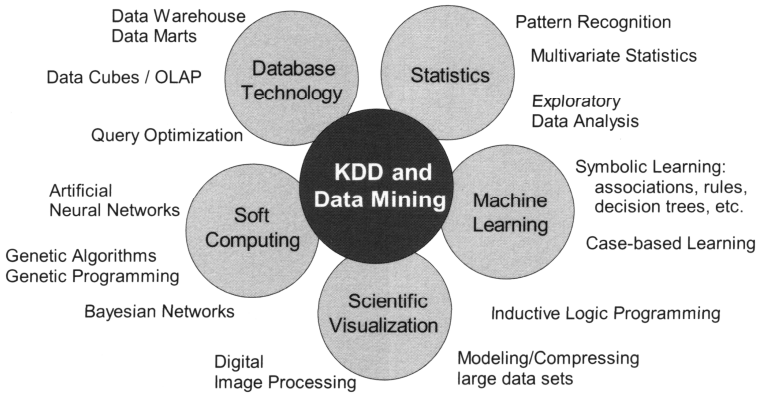


Fig. 3: KDD and Data Mining: intersection of multiple disciplines (from [Otte04], p. 22)

Some of the central terms of the related research are “knowledge”, “domain knowledge” and “application domain” as well as their relations. Further on, the following definition will be used according to the definitions from Manoel Mendonca and Nancy L. Sunderhaft [Mend99], p. 7:

> *[a]* “Domain knowledge is non-trivial and useful, empirical information specific to the application domain believed to be true by the data users. Background knowledge is the domain knowledge that data users had before analysing the data. And, or discovered knowledge is the new domain knowledge that data users gain by analysing the data. Domain experts are data users that have a sizeable amount of expertise in a specific application domain.”

From the humanities point of view, dealing with knowledge always has something to do with the process of acquiring, understanding and sharing new things. Here the perspective of a business informatics researcher will be the master perspective. From that perspective, the new knowledge is what can be described (verbal or formal), perhaps firstly unknown and is generated out of data or information, but it can – in all different aggregates in which it can occur – be computed and processed.

Known domain knowledge can be documented by the use of formal languages and methods, e.g., ontology’s [Abec04] or description logics

[Baad04]. That is possible for explicit knowledge. Something is known and its semantics is coded according to a certain description language.

The other focus of this text will be ‘knowledge generation’. Digging for new knowledge in data has commonly been called data mining (DM) since the early 1990s [Mend00], p. 6:

> *[b]* “Data mining can be defined as the process of extracting new, non-trivial and useful information from databases.”

One of the most popular and commonly accepted description of the data-mining process (DMP) was created by U.M. Fayyad, G. Piatetsky-Shapiro, G. and P. Smyth in their paper “From data mining to knowledge discovery: an overview”, published in 1996. They describe the data-mining process as follows:

> *[c]* “KDD is the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data.”

KDD and Data Mining will be used synonymously in this text, whereas some researchers make distinctions between these two terms.

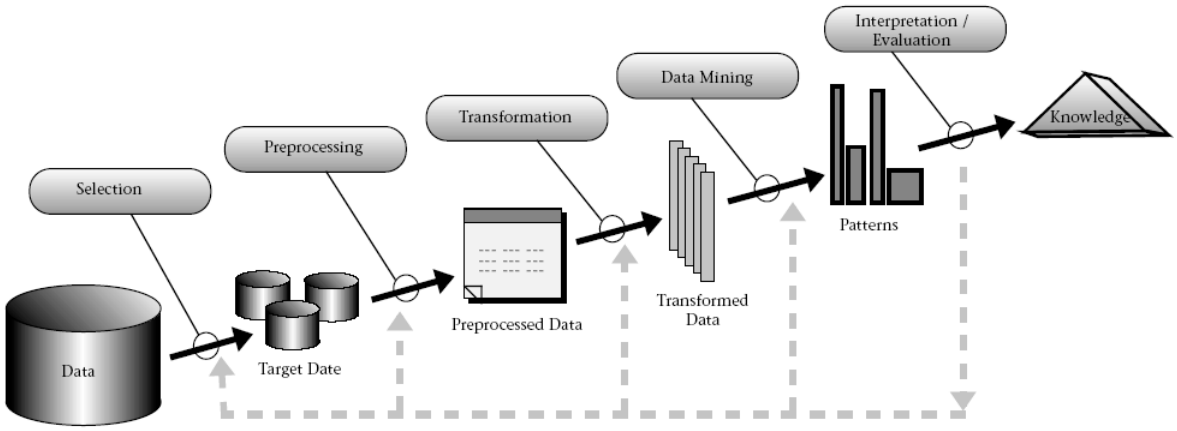


Fig. 4: Data-mining process (from [Fayy96], p. 10)

Fayyad et al. described five main data-mining steps (see Fig. 4). In the *Data Selection* step a subset from a set of data (a set of facts, e.g., cases in a database) is selected for further *Target Data Pre-processing*. This step is necessary because data selected during the first step is usually not ready for data mining. It can occur in several data formats which have to be transformed into a common format. Then the *Pre-processed Data must be transformed* into a data-mining-prepared tabular format in which the columns represent the so-called “features” (objects, which are observed during the DMP) and the rows that store all occurring values. The *Data Mining* step itself is processed in this prepared data and can be based on several main techniques:

- Classification Trees
- Association Discovery Techniques
- Clustering Techniques
- Artificial Neural Networks
- Optimised Set Reduction
- Bayesian Belief Networks
- Visualization and Visual Data Mining

The goal of this text is not to be an introduction to data-mining methods. Therefore, for details one is referred to standard books that give a good overview on data mining[8].

Probably the most important step within the DMP is to evaluate and interpret the patterns found. Here the conversion of patterns to knowledge is done, which is crucial for the success of the data-mining process.

From the knowledge management point of view a transformation of raw data takes place that generates wisdom that consists of knowledge and experi-

[8] E.g.: Hippner, H.; Küsters, U.; Meyer, M.; Wilde, K.D.: Handbuch Data Mining im Marketing, Vieweg, 2000 or Brebbia, C.; Ebecken, N. F. F.; Zanasi, A.: DATA MINING VI - DATA MINING TEXT MINING AND THEIR BUSINESS APPLICATIONS, WIT Press, 2005

ence through the transformation of data into information by adding context and rules that turn it into knowledge (see Fig. 5).

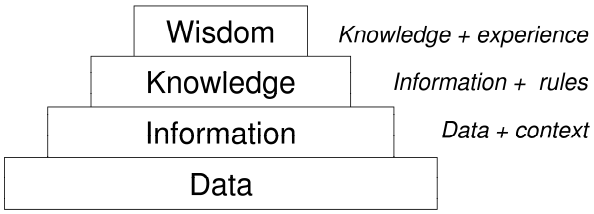


Fig. 5: Data Pyramid (from [Grob03])

Later on, with regard to efficiency, it must be ensured that knowledge that was discovered in the past is used as a basis for the adjustment of current decisions, which may be relevant for the future. That means that the DMP must be extended by a learning component which permits the consideration of past results.

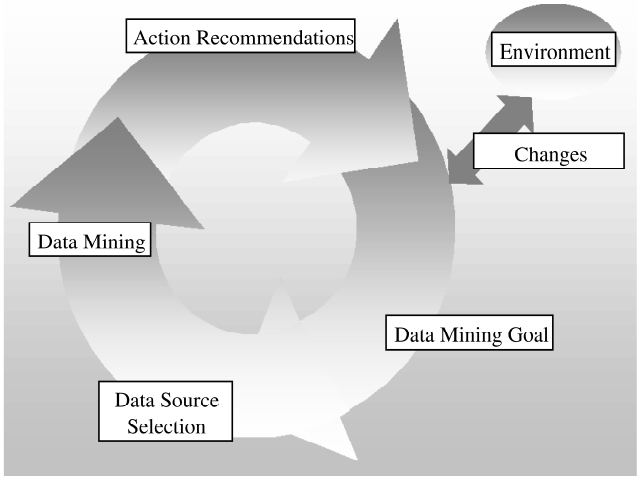


Fig. 6: Data-Mining Management Cycle, translated from [Kall04], p. 52

To do this, a Data-Mining Management Cycle has been proposed in [Kall04] (see Fig. 6).

For textual data the DMP was adapted according to the specialties of this kind of source by several research activities. Hidalgo [Hida02] proposes an adoption of the KDD Process for textual sources in a way that every single process step is assigned to usual activities in language processing (see Fig. 7):

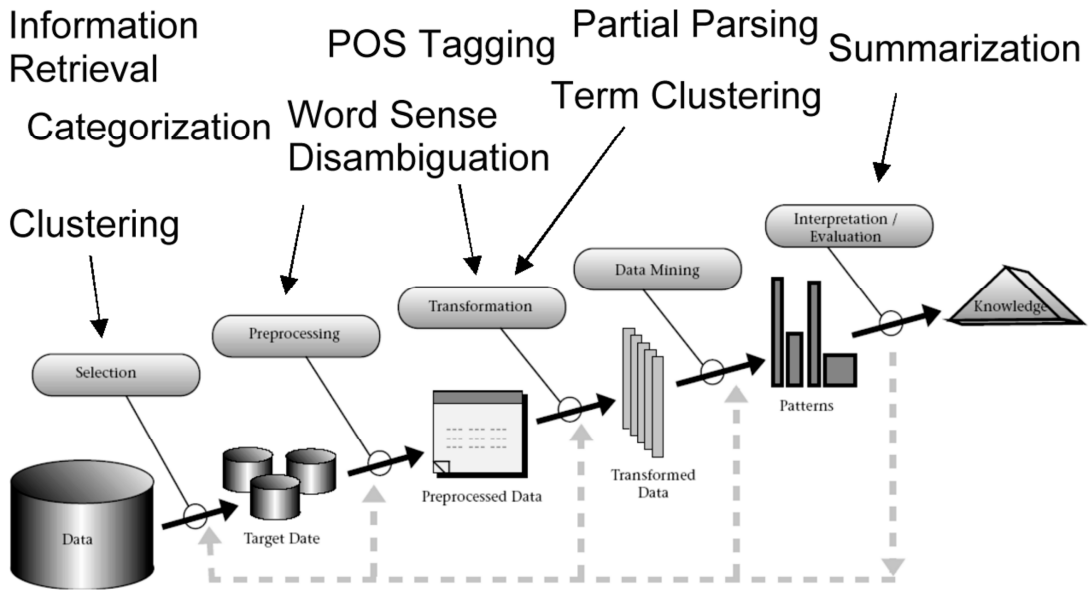


Fig. 7: Text Data Mining adapted to the KDD Process steps (adopted from [Hida02], p. 9)

This view is only one of several possible perspectives on textual data mining. Alternatives will be discussed when the method that is proposed in this text is introduced. All the processing, which is done on the data, must be appropriate for the aim of research. Therefore, in the following, the alternatives, appropriate process steps and methods have to be evaluated, which allow for the observation of developments in knowledge domains using large textual sources.

In this chapter the main concepts and objects will be introduced and made subject to a definition where the development of further methods will be based on a common understanding.

2.1 What is a text?

Due to the focus on textual knowledge representation, the source on which the knowledge extraction process will be applied has to be clarified. Different research disciplines use different definitions for a “text”. Here the computational linguistics view on texts is used. “Text” is used in its limited form – a domain-related text collection. Wikipedia, the independent free online encyclopaedia project gives an orientation of the wide semantics of the term “text” [Wiki06]:

“The term "text" has multiple meanings depending on its context of use:

- In language, text is a broad term for something that contains words to express something.
- In linguistics a text is a communicative act, fulfilling the seven constitutive and the three regulative principles of textuality. Both speech and written language, or language in other media can be seen as a text within linguistics.
- In literary theory a text is the object being studied, whether it is a novel, a poem, a film, an advertisement, or anything else with a semiotic component. The broad use of the term derives from the rise of semiotics in the 1960s and was solidified by the later cultural studies of the 1980s, which brought a corresponding broadening of what it was one could talk about when talking about literature.
- In mobile phone communication, a text (or text message) is a short digital message between devices, typically using SMS (short message service). The act of sending such a message is commonly referred to as texting.
- In computing, text refers to character data or to one of the segments of a program in memory.”

In CL, elements of language are subjects of research that may consist of one or more language element. These various elements will not be defined here in particular, but the object of research must be made clear. CL researchers use “Term” and “Word” as a name for language elements between two spaces in basically the same way. A “Term” is commonly used for n-grams of words (2 or more words). Here “Term” will be used more generally. “Term” was previously used intuitively; it will now be defined as follows:

> *[d]* “A Term is an element of a text that consists of alphanumeric sign combinations, which occur between two blanks. A Term can be constituted by one single term or more terms. It may begin with a number concatenated by space or beginning with a letter concatenated by a minus “-“ or a point “.”.”

This wide definition covers product names and technical norms, which are very important for the later analysis of the trends in technical domain corpora. As terms are also assumed technical norms (e.g., X.25) or names of companies (e.g., abbreviation: ITT, compound word: Hewlett-Packard).

When dealing with sentences these terms are used in a linguistic manner as texts:

> *[e]* “A text is a number of grammatically correct grouped terms, used to describe facts and theories within a certain knowledge domain.”

Within the additional work the term “text” is understood in the clarified meaning of definition 0.

> *[f]* “A text is written language that is semantically related to a certain or more knowledge domains. It is potentially convertible into a computer-readable format, e.g., ASCII.”

If tracking semantics of knowledge domains is in focus, it has to be decided which elements within a text may indicate a semantic trend. Further on, the term “text” will be used for any kind of textual source data in different formats, whereas a “corpus” is mentioned as pre-processed textual data.

> *[g]* “A corpus *C* is pre-processed textual data in the context of text analysis or TDM.”

Additionally a corpus in CL is usually described by corpus metadata, at least: Author, source, and date. A differentiation and clarification of texts within this work will be done in Chapter 3.2.4.1.

2.2 What are “Trends” and “Hypes”?

Here I do not want to join in the discussions that social trend researchers are conducting. A social trend in this meaning is something not easy to describe – a development that leads to changing market behaviour of customers. Therefore, this kind of trend is analysed by several investigators, some of them with mystic auras and, on the other hand, very involved in the context of market research. One example is the Gartner Group that introduced their “Hype Cycle of Emerging Technology” (see Fig. 8).

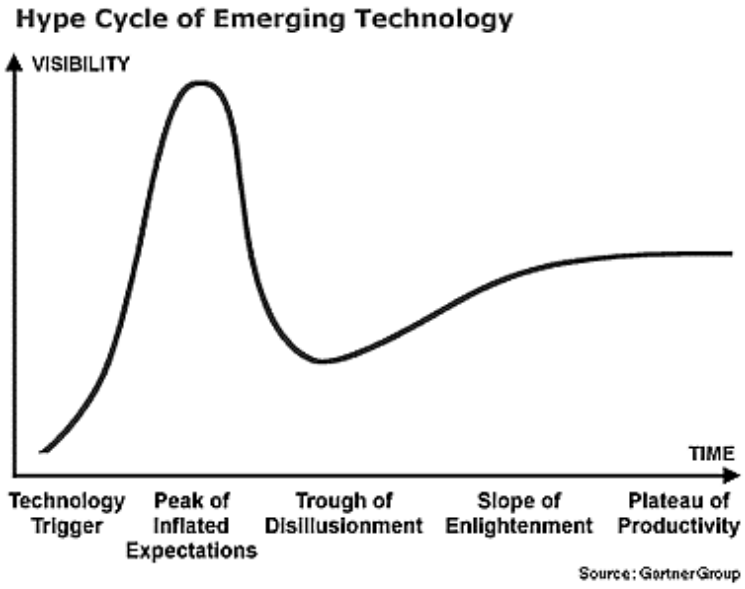


Fig. 8: Hype Cycle of Emerging Technology (source: [Gart03], p. 5)

Gartner’s Hype Cycle became very popular in recent years and every new technology is actually assigned to this curve according to Gartner’s estimation of their progress status.

The five separate phases in the progress of a technology are:

1. Technology trigger – a breakthrough, public demonstration, product launch or other event that generates significant press and industry interest.
2. Peak of inflated expectations – a phase of over enthusiasm and unrealistic projections during which a flurry of publicized activity by technology leaders results in some successes but more failures as the technology is pushed to its limits. The only enterprises making money at this stage are conference organizers and magazine publishers.
3. Trough of disillusionment – the point at which the technology becomes unfashionable and the press abandons the topic because the technology did not live up to its over-inflated expectations.
4. Slope of enlightenment – focused experimentation and solid hard work by an increasingly diverse range of organizations lead to a true understanding of the technology's applicability, risks and benefits. Commercial off-the-shelf methodologies and tools become available to ease the development process.
5. Plateau of productivity – the real-world benefits of the technology are demonstrated and accepted. Tools and methodologies are increasingly stable as they enter their second and third generation. The final height of the plateau varies according to whether the technology is broadly applicable or only benefits a niche market.

Equal phase models were known in market research before but the popularity of the Gartner model is very great. Here a pragmatic understanding will be used to deal with the term trend that is more compatible with statistical understanding. As an extension to the idealized progress courses proposed by Mertens [Mert95], p. 25 the following distinction is made between developments (according to their persistence characteristics):

> *[h]* “A Trend is a steady growing or falling occurrence of the same semantic concept within a certain domain that influences the progress of the domain in the long run.”

In general a trend can consist of one of four possible qualities: The underlying topic may become more popular or less popular. Perhaps a trend can also decline over time or remain constant. In the last case a pure “Trend” is not what is meant, but rather this constant concept as a special kind of trend which will also be observed. Due to the non-constant nature of popular measures a threshold must be defined to approximate “constants” in domain progress.

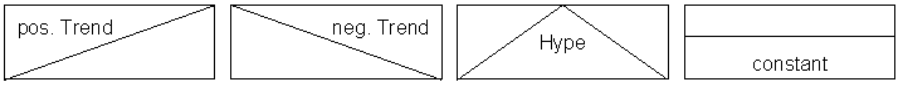


Fig. 9: General trend shapes

In [Kont04], p. 204ff the “Patent Miner” application is introduced, which uses a kind of shape description language to process and retrieve concepts according to the different trend curves introduced.

Another special kind of trend is a “Hype” that is defined here as follows:

> *[i]* “Hype is a semantic concept within a certain domain which only dominates a short period of time.”

Hypes are rather "nine-day wonders" after this interpretation, without considerable durable meaning. “Trends”, on the other hand, characterize the domain in the long run.

After defining this basic terminology, possible measure candidates will be discussed in the next chapter which will then be made operational for a corpus-measure-based observation of domain progress.

2.3 Challenges with progress extraction from text approaches

The amount of text may led to serious run time performance issues, especially if the source data belongs to a living domain where texts are perma-

nently added and not to a restricted archive that is organized for information retrieval only. This is especially the case when no sequential processing is applied to the data and the whole knowledge must then be completely stored in the memory during run time. In spite of stop-word filtering, there is no other generally applicable pre-filtering approach available. This leads to a high dimensionality in data and results that are biased by noisy, not-domain-specific data. The crowd problem leads to the general knowledge extraction challenge: How to find aggregated, potentially interesting patterns, which reflect the semantics of the text source properly.

Most of the methods used analyse text corpora word by word or sentence by sentence and use clustering and tagging techniques (for a comprehensive approach see [Spil02]). Usually the whole data set is taken for further analysis. The methods work “bottom up”: Taking each data element, and computing for interesting patterns.

Other methods are more statistically based and use Vector Space Models which convert the whole text corpus into a vector representation of occurring terms, see [Dame05] for an introduction of approaches. They do not explicitly consider this (only partly by using stop-word lists) if the whole data is mixed with a high level of noisy data (text with no technical domain-specific or time-dependent information) which may negatively influence the precision of further analysis procedures.

Classical Text Mining approaches that convert this data into a computable format have to consider the specifics of texts, e.g., incomplete matching thesauri and sparse problems, which aggravate the use of methods from the data mining research field. Before such DM methods can be used, the data has to be pre-processed in a way that the resulting data set fits the requirements of the DM methods used. Every mismatched text element that is not worth later analysis, but remains in the process, lowers the quality of the found knowledge that is extracted out of the corpus.

Romanov et al. [Roma05] introduced the knowledge discovery approach that dynamically changes the thesaurus according to the structures in large text databases using the MST algorithm. The MST algorithm is based on similar-

ity measures that allow relation recognition between terms. The dynamic here comes from adjustable pair frequency of terms that have adjustable weights for covering changes. Thus, the method allows the cutting of weak links between terms. Due to computing difficulties only 2000 database records and 3000-5000 terms are executed simultaneously.

The current research in TDM and linguistic computing is very fragmented. This circumstance exposes the danger of reinventing methods by researchers who are unfamiliar with the whole canon of TDM methods from all related research fields. Reasons for that are probably the gaps that exist in research objectives and methods between humanities, computational linguistics and informatics. Where the first group struggles with the question of whether quantitative methods are appropriate, the last group of researchers sometimes seems to have problems with some quantitative measures that are inappropriate or which need special preconditions in use for certain tasks.

That is especially true for trend extraction approaches that deeply rely on quantitative and statistical methods with reliable thresholds as a basis for precise decisions whether a concept belongs to a trend or not.

But this dissertation cannot prevent all such issues. A claim to know all specific solutions in DM will not be made here. By focusing on the preprocessing in the task of tracking domain-specific progress and by providing a framework that is complementary to known DM methods this risk is minimized.

Known common DM problems with textual data based on the KDD approach are (see [Fayy96], p. 26):

- Large (textual) data collections
- High dimensionality
- Over fitting
- Changing data and knowledge
- Noisy data
- Understand ability of mined patterns

Hidalgo ([Hida02], p. 7) expressed some more problems:

- Text is not designed to be used by computers
- Complex and poorly defined structure and semantics
- But much harder, ambiguity
  - In speech, morphology, syntax, semantics, pragmatics
  - For instance, internationality
- Multilingualism
  - Lack of reliable and general translation tools

Surprisingly the majority of publications in TDM do not consider pre-processing qualities of text sources and take a text collection “as is”. Even when a dependency of TDM results from the quality of text collections is to be expected, no sufficient empirical research on quality aspects or quality measures for certain TDM tasks are currently known.

2.4 Implications on current work

Overcoming the lack of research in handling large data sources needs to extend pre-processing of source data in a kind that allows considering contents of whole source data in later processing without computing whole data together at a time.

The proposed approach here uses corpus segmentation that is oriented to the semantic of the domain and is also capable of dynamic adjustment while clustering new terms according to their persistence (constant or volatile). It can also be seen as a method for automatic stop-word generation. Stop words in this meaning are terms that have only low information value because they persistently occur over longer time periods. Opposite to the persistent terms, the other cluster contains volatile terms that occur very dynamically within the corpus. According to the results of Zipf (see [Zipf49]) it is to be expected that the constant and the volatile terms have significant different statistical qualities that allow dividing them into separate clusters and apply tailored methods to both clusters in following DM processes.

3 Domain progress extraction

In this chapter the task-specific approach of tracking domain knowledge based on segmented textual sources will be explained in detail. For this the trend-mining framework (TMF) will be introduced that consists of process phases that were adopted from Fig. 4 and precised according to the aim of research, on the one hand, and tools for supporting the mining process, on the other. The aim of the development of the TMF was the specific task of processing a large collection of textual data with the aim of progress extraction of a certain domain. Thus, it is assumed that the TMF process starts after the task of TDM is described and the data source selection will have already taken place. The TMF process starts with the pre-processing step.

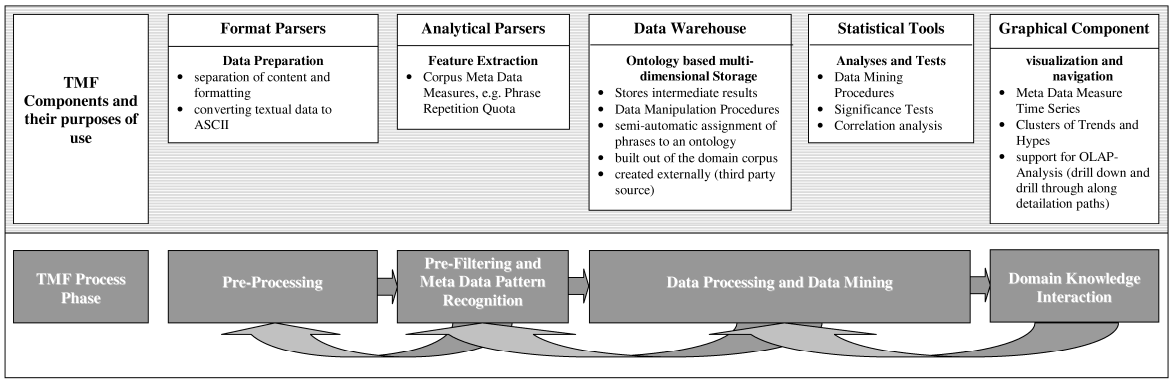


Fig. 10: Overview of components of the TMF and their use in the Trend Mining Process

Fig. 10 shows both perspectives: The TMF components and their use in the TMF process. The TMF is a proposed methodology and also a process of TDM which makes it possible to extract time-related domain knowledge based on unstructured textual data semi-automatically. The main goal of developing the TMF approach is to establish a framework that permits exploring and analysing large Technical Domain Corpora in an intuitively interactive way. For this, methods were evaluated which allow measuring the quantitative characteristics of a time tracked domain corpus.

The TMF consists of

1. Format Parsers: For separation of content and formatting as well as converters to ASCII
2. Analytical Parsers: For determination of the corpus measures and Feature Extraction, e.g., TRQ
3. Data Warehouse: For the storage, e.g., of intermediate results and for a semi-automatic assignment of terms to an ontology out of the domain corpus
4. Statistical Tools: For analyses and tests
5. Graphical component: For visualization and navigation

The main specifics of the proposed methodology are:

- Applying a horizontal and vertical segmentation on source data
- Using corpus measures and defining thresholds for pre-filtering terms
- Aggregating terms to concepts with the use of domain-specific taxonomies
- Visualizing extracted concepts in an OLAP-based intuitive navigation concept

In the TMF process the main task of automatic extraction, aggregation and interaction with extracted domain knowledge is divided into a few sub-tasks. The tasks are performed by the different components mentioned earlier. In its conceptual flow this chapter follows the common TDM process (see Fig. 4 and Fig. 10) with the following steps:

- *Text source selection*
- *Pre-processing*
- *Pre-filtering and corpus measure pattern recognition*
- *Data processing and data mining*
- *Domain knowledge interaction*

In the next chapters the above-mentioned process steps within the TMF will be introduced.

3.1 Text source selection

The domain knowledge that is to be tracked is only represented by textual sources in this approach. Because real corpora are finite, the explicit knowledge about the domain is also limited by the used corpus. Senellart et al. [Sene04], p. 26 assume in their research that a thesaurus of a corpus is domain specific to this corpus. Therefore, better results in domain-related research are to be expected the larger and more domain related a corpus is. In order to found the empirical analysis solidly, a procedure has to be compiled so that the necessary width can be taken from the procured source. However, a detailed depth permission for the processing of the data selection is needed which makes an investigation possible on term level. To illustrate the economical framework above all the following criteria for the selection of a text source are of high importance:

- Semantic constant of the time series definition during a long period (avoidance of breaks)
- Existence of sufficiently long data acquisitions
- Access and evaluation possibility under consideration of economic criteria
- Authenticity of the data

Due to these targets above, it is possible to use all data sources which are available in a standardized form or available over longer periods from service providers of high reputation. Therefore, in particular, business reports and end-of-year procedures of enterprises are applicable, since their production is based on underlying regulations (according to HGB, the law for companies in Germany) and are essentially stable over many years. Unfortunately these data sources exhibit a very high aggregation degree of data which makes conclusions not as adequately possible as those initially viewed articles of designated elements of the IT or the operational reality. Register of companies entries show a high degree of standardization, but they are, however, only suitable for special questions [Spil02], pp. 117 due to missing depth of detail. Therefore, the end-of-year-procedure attached reports of management offer a better starting point. However, less restrictive regulations are applied

regarding the degree of detail, which make comparability particularly more difficult between substantially different enterprises. Information from such sources should therefore be considered only as an addition in the framework of case examples or in connection with other data use. Data collections of the federal statistic office or by trade associations are more suitable, whereby it must also be considered which data were collected, being ex-ante explicitly defined. This means that current developments will possibly not be considered or, if they are, then only in the future.

An attempt to overcome the aforementioned difficulties is through the use of historical sources, which were not revised in between and thus potentially reflect the former reality. Such unstructured data as mentioned escape, however, an immediate computer-aided evaluation. Manual approaches can only provide restricted results in economically acceptable time periods but offer, however, valuable clues about further-reaching analyses [Mert95], pp. 25. Special suitable text documents on hand, which can be won from different source formats as extracts, can be considered in electronic form (plain text, e.g., ASCII). Zanasi [Zana05b] discusses various problems that occur when using information in the public domain – mainly a so-called “Information Overload Problem” and the “Specialistic Language” that may represent knowledge in sometimes hidden levels.

The challenges from the inhomogeneity of text sources and the differences in knowledge representation are the elimination, or at least the consideration, of this inhomogeneity by the methods applied to avoid biasing of results.

3.1.1 Methods to exploit domain knowledge in the mining process

In linguistics all contents of corpora are mainly converted into a vertical representation[9]. Then specific linguistic processing is applied, e.g., stemming and POS tagging. Linguistic approaches do not fit if the focus is not the language itself, but the semantic of discourse that is represented by the text. In

[9] Meant here is the logical approach, not the physical implementation. Normally a text remains readable after this conversion process.

technical domains abbreviations, proper names et cetera represent important parts of domain semantics. But this special quality may not be properly covered by linguistic methods that focus on the language only.

There are many tools available on the market, which are used to support knowledge discovery. In [Kühn02], p. 1 an evaluation of various tools that play an important role on the market for text (data) mining is described. They can be categorized in five classes of systems:

1. Systems for text exploration and extended text search
2. Systems for text and data mining
3. Development tools for text (data) mining
4. Systems for answer extraction
5. Systems for content analysis

The tools are sophisticated to different extents and deliver different technologies and functions in the fields of input interfaces, categorizing, answer extraction, data mining with pre-defined or flexible language support.

For the text analyst, using these built-in technologies means partly giving up the freedom of how to decide how the data is to be handled. Tools always use semantics which are pre-defined by the tool vendor. The tool as a “black box” provides only a couple of parameters that can be adjusted by the user of the tool.

Typical methods implemented in tools for knowledge discovery are clustering, categorizing and naming entity recognition among others. Some of them are based on thesauri or stop-word lists that are pre-defined (most of the tools permit extending these lists).

What does it mean to the analyst? The analyst may use these lists and pre-defined categories to work on his data material. He will get a categorized and clustered output. Nevertheless, how the output is generated in detail cannot be fully understood. The investigation is partly given to a tool that is not completely under the command of the analyst.

What does it mean for the results of the analysis? One of the main functions of knowledge discovery tools is to support information retrieval. For that reason algorithms are implemented, which try to categorize texts in categories like “bill”, “invoice”, “customer inquiry”, for example, among others, for the purpose of pre-sorting incoming letters for further action.

The view on data, which is applied by many tools, is a timeless view. This implicates an assumption that all relevant semantics exist in an even distribution throughout the input data. To understand why the time dimension is not considered in many cases, it is necessary to think about the difficulties of extracting time information from a text sample. To find a date may be easy by using pattern-matching technology and looking for the typical format(s) of date statements. To extract dates from the text or to get this information out of the text (-file)-corpus measure is not a problem from the technical point of view. Nevertheless, to categorize a text regarding the date of generation or according to the time which the text content documents, is not easy to compute. Information regarding this is for the most part extracted by analysing the semantics or the context in which the text is written. The choice of using a special methodology should always be guided by the goal of the analysis. The user has to decide which level of discrimination is needed in the results. He always has to keep in mind that there is a technical limitation determined by the methods, which cannot be overcome when using that special tool. The gap is that only a few parameters can be adjusted to fulfil the analyst’s requirements. It is very helpful to make use of automatic algorithms for pre-sorting masses of written texts or a first overview about a topic of interest. Another question to consider is for which process step of the analysis the support of a tool is recommended.

Analysis software usually uses thesauri or stop-word lists to classify whole texts or single words. Indeed built-in algorithms have limitations. That might be a problem if the result cannot be reconciled because the internal logic is not known. The effect of the application of taxonomy will be the focus in a later chapter. The other problem is that which the built-in logic has in distinctions between the allocation of words or texts based on basic rules, especially when semantics have an important meaning. One example may illus-

trate this issue: To allocate words by using basic linguistic rules like allocation by word trunk does not produce meaningful results in any constellations. One German example should illustrate which wrong conclusions can result when simple rules are used: "rechnen", "Rechner", "Rechnung" or in English: "calculate", "calculator" and „bill". From the German linguistic point of view the three words have the same word trunk but the semantic meaning from the information-technological point of view is different. In the author's opinion, the analysis is biased before it is started. Therefore, the results will also be biased and the goal of extracting new and valid knowledge will not be reached on a high quality level.

Other examples of this gap of pure linguistic methods are the words "interessant" (in English: "interesting", may be anything) and "Interesse" (somebody can be interested in something). But an "Interessent" (in English: "prospective buyer") should be interpreted as a possible future customer. From this analysis view, there is equivalence between "Kunde" and "Interessent". This shows that there is no linguistic link (in the German language) rather a semantic link between these three terms.

How is a "word" or a "term" defined in the implemented logic of the tool? That is one of the most important questions. Investigating the context of texts regarding business informatics domain, it is not the same to analyse plain text written by a poet or to analyse a novel. IT vocabulary contains terms that are normally not assumed as "words" like technical norms (e.g., X.25) or names of companies (e.g., abbreviation: ITT, compound word: Hewlett-Packard). An additional important point is the ability to handle language-specific characteristics like the German "Umlaut" and other non-ASCII codes.

The aspect to be able to define which terms are analysed and which not is fundamental for the results given by the analysis. Standard tools that fit all the requirements are rare. Some common approaches and their implementations in existing tools are introduced by Zanasi [Zana05c].

Due to limitations in the configuration of "standard" tools, the TMF used here was established as a "best of breed" approach that was preferred instead of losing methodological control by using only one tool that is limited in one or

methodological aspects. The advantage here is to keep full transparency throughout the whole process of knowledge acquisition.

3.1.2 Excurse: Underpinnings of evolution in business informatics magazine titles

The development of the terms and contents belonging to the business informatics domain is reflected in (German) literature, in particular in the titles and main topics of specialized publications. At the beginning of the 1960s, the term “technology” would refer to computing technology, cybernetics or electronic data processing speech, whereas today computer-assisted data processing would be understood. Exemplarily the specialized publications “Elektronische Informationsverarbeitung und Kybernetik: EIK” (1965-1986) and their follow-up publications “EIK: Journal of information processing and cybernetics” (1987-1991) are mentioned. As technical developments of these epochs balancer, multiplier and later electronic computers were mentioned. At this time the term “electronic data processing” or EDP was developed, which is still used today. A different content-wise emphasis was to be observed due to the technological level of development (see the appearance period of the magazine “Elektronische Datenverarbeitung” (10/1968-12/1970).

Computing technology was first predominantly an engineering challenge, up to the point that one regarded it under the perspective of its later use (see magazine “Elektrotechnik und Maschinenbau: E und M” (1967-1987), “Elektrotechnik und Informationstechnik” (1988-) developed in the course of the 1970s and 1980s more and more specialized in several directions. One example of the focusing on certain sub-ranges of the electronic data processing is the “Elektrotechnische Zeitschrift: ETZ” (1922-5/1995 with interruptions). The title of this magazine changed again: “ETZ: Elektrotechnik+Automation”. This properly reflects the change in technological progress. With this, the development of a self-understanding of the business informatics was connected as supporting technology of most different functional areas of the operational

reality (an example is mentioned: “Industrielle Informationstechnik: IT/AV” (1998 -)).

Within the production sector IT firstly supported areas close to operations. Technical equipment that was initially named “control engineering”, was later called “process computing technology” and, afterwards, “automatic control engineering”. Whereby during these terminological changes surely content-wise emphasis shifts played a role. On the other hand, it has to be accepted due to the content-wise relationship of the topic areas that also fashionable aspects may have been a factor in these new terms being coined. Parallel to this application orientated view emphasis, it can also be observed that the theoretical-scientific view focus was always present and has continued to develop. Due to that, in the course of time, various special disciplines developed a variety of application and effect scenarios of business informatics based on the initial mathematical-cybernetic adjustment. Examples to mention are: Information techniques and business informatics, computer, systems, and applications. Due to this, in 1986 the specialized publication “Elektronische Rechenanlagen” was renamed “Informationstechnik: it; Computer, Systeme, Anwendungen” and in 1993 a successor received the title “Informationstechnik und Technische Informatik: it + ti”. At the same time an appropriate content-wise emphasis shift was carried out.

3.1.3 Text Source 1: WWW Archive of German “Computerwoche”

The German weekly magazine “Computerwoche”, from now on abbreviated as “CW”, has continuously been published since first being featured in 1974. In 1974, only five issues were released. Since 1975, a weekly feature cycle has been established.

With the available past experience concerning the editorial high quality and the knowledge about the regularly reflected width in contents of the CW, the CW is evaluated as a medium that comprehensively reflects the operationally relevant business informatics domain facts during the observation period That’s why this publication is suitable for an investigation of information-technology development. In the timeline of downloads starting in December

2002 and ending in March 2004 a total of 152,283 links to different CW articles were found in the archive for the period from 10/1974 to 12/2003. From this source 132,406 articles (a share of approx. 87% of the total number of articles) were downloaded and evaluated. In order to ensure a high sample quality, for each year the share of downloaded articles was computed. As minimum share 66% were defined. For years with lower values one or more supplementing loading procedures were accomplished. That was the case for the classes 1976, 1977, 1982, 1985, 1986, 1987, 1988, 1989 and 2000. This was caused by technical limitations. Due to distributed data retention of the articles on four Web servers, partly not functioning forwarding links (“Click here”, “Page has moved”), dead links (“This page is currently not available”) were found or at the Web server maintenance work took place. In the result completeness ratios resulted between 66.21% (1994) and 100% (1975). It has to be assumed, due to the length of the load time area and the made reloading processes, the unavailability of certain articles is evenly distributed stochastically over all classes and expenditures of the CW. A systematic influence of the sample regarding feature classes or certain contents can therefore be excluded.

The first year of issues was not considered due to a lack of comparability: The issues started in October 1974 and contained only 229 articles, whereas during other years they contained up to 5,000 articles. Because of extreme corpus length differences and their influence on dependent measures (see [Baay98]) the year 1974 was judged not to be comparable to the other years and therefore excluded from further analysis.

The design of the publication regarding content is pointed out to be influenced by changing main focuses of the editorial staff in [Mert95], p. 28. There is a good chance that journalistic emphases have been carried out in the course of time. This circumstance is not considered problematic, however, because this development is an expression of the reflection of the information technological domain reality of the respective epoch. If the reporting focus of CW had stayed at the level of mainframe machines, while ignoring the latest technical trends, then CW would not have been a proper source for this dissertation.

A thematic preference of authors is another aspect. The fact has to be accepted that it will not produce a one-sided distortion due to the multiplicity of authors over period of observation but rather a reinforcement effect is to be expected, which concerns mode topics, because with such topics like this no author would like to remain “outside forwards”. All in all, this publication is rated as suitable for the examination of the information technological development.

3.1.3.1 Semantic benchmark for text source 1

Which results are to be expected when analysing a large business informatics related online archive? In commercial or governmental research projects like TDT the expected results are pre-coded into the official test sets of texts. In this real-world scenario more general expectations must now be formulated, because no ex-ante knowledge regarding the text contents is available (without analysing the text). Looking back to Mertens [Mert95], who manually analysed the same source more than 10 years ago, it is to be expected that the main results he found on the semantic level should also be extracted with the methods applied here. It must be considered that Mertens and his team analysed the CW article by article, and assigned each of them to roughly predefined categories that were extended during the analysis process. From this process a number of graphs were produced that documented progress paths over the observed time period. Mertens then summarized the headwords for each year and labelled them with a predominant topic (see Fig. 11).

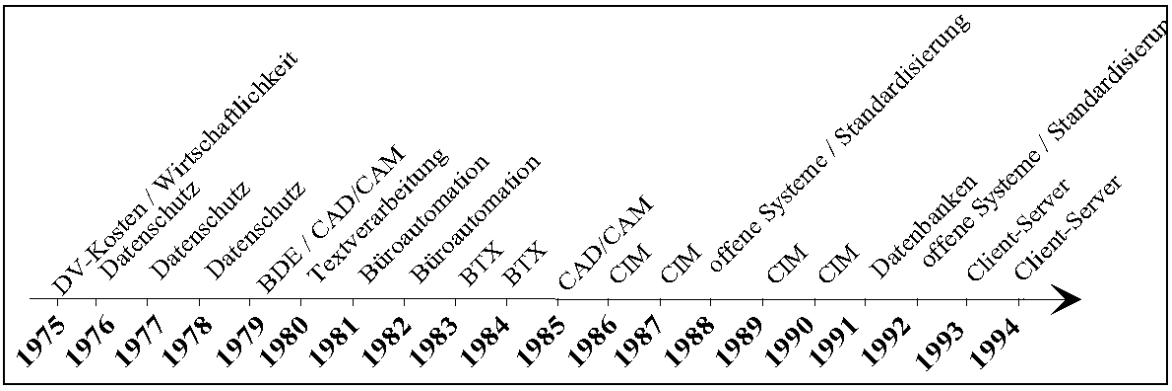


Fig. 11: In the course of time from 1975 to 1994 respectively most frequently mentioned business informatics domain-related headwords (see [Mert95], p. 32)

The applied methods used in the current process of TDM will be judged using Fig. 11 as benchmark. Due to differences in analysis methods, the semantic similarity, not the exact term matching will be compared. It is also to be considered that due to the limited amount of analysed issues of the CW and the different methods applied, the extracted results by Mertens are not a real benchmark and may differ due to differences in the extraction process (manual vs. automatic, complete articles vs. headwords).

3.1.4 Text Source 2: Printed Allianz Management Reports

As another application for a real-world evaluation task, the analysis of a much focused publication seemed to be appropriate. Management reports are potentially suitable for knowledge extraction due to their relatively persistent structure. The source introduced here is a source of printed Allianz management reports (“AI1k” or “AI100”). Compared to CW, this source has a minor quantity of only a thousand tokens per year.

1962 *German*	1963 *German*	1964 *German*	1965 *German*	1966	1967	1968 *German*	1969	1970 *German*	1971 *German*
1972 *German*	1973 *German*	1974 *German*	1975 *German*	1976	1977 *German*	1978 *German*	1979	1980 *German*	1981 *English*
1982 *German*	1983 *German*	1984	1985 *German*	1986 *English*	1987 *English*	1988 *German*	1989 *German*	1990 *English*	1991 *German*
1992 *German*	1993 *English*	1994 *English*	1995 *German*	1996 *German*	1997	1998 *English*	1999 *German*	2000 *English*	

Fig. 12: Considered Allianz management reports with language of origin (missing reports marked grey)

Seven reports were not available. They were marked grey in Fig. 12. Others were only available in English. The financial reports were explicitly not considered, but all written text within the management reports, including special (or featured) topics that were added in later issues. All reports were translated into German with the use of a computer translation program and the application of an automatic spelling check to avoid serious grammatical and writing mistakes that might have influenced later processing. By so doing, a manual bias should have been avoided. This processing must be considered in the following analysis and interpretation, because influences from automatic processing indeed may bias later results. This source represents a non-optimal, but real scenario for a check of robustness of results found.

3.1.4.1 Semantic benchmark for text source 2

Allianz is a well-known internationally active insurance company with a long history. Many facts were collected over time. Also Allianz' public relations department's press releases do reflect the company. The following facts that have major importance within the developmental history of Allianz are taken from their Internet site (see Fig. 13):

2000	Purchase of the US asset management company Pimco Advisors L.P., Newport Beach, California First listing of Allianz stock on the New York Stock Exchange
1999	Beginning of expansion in Asia: e.g. founding of a joint venture in China, Allianz Dazhong Life, and acquisition of First Life Insurance Co., Ltd, South Korea Introduction of the new Group logo for all Group companies
1998	Establishment of asset management as a core business activity through the creation of Allianz Asset Management, Munich
1997	Development of Vereinte Krankenversicherung AG into the health insurance provider of Allianz Acquisition of Assurances Générales de France (AGF), Paris
1995	Acquisition of the ELVIA Group, Zurich, Lloyd Adriatico, Trieste, and the Vereinte Group Purchase of a stake in the Australian insurance provider Manufacturers' Mutual Insurance Group, Sydney (today Allianz Australia Limited)
1991	Acquisition of the US insurer Fireman's Fund Insurance Company, Novato, California
1990	Takeover of the state insurance company of the former German Democratic Republic (East Germany) Beginning of activities in Eastern Europe, e.g. purchase of Hungária Biztosító, Budapest
1989	Purchase of a stake in the French insurance group Via/Rhin et Moselle (today a part of AGF)
1986	Takeover of Cornhill Insurance PLC, London
1985	Formation of Allianz AG as a holding company
1984	Purchase of a stake in Riunione Adriatica di Sicurtà (RAS), Milan
1976	Establishment of the property and casualty insurance business in the U.S.
1974	Foreign expansion stepped up: including Great Britian, the Netherlands, Spain, and Brazil
1966	Opening of a management office for Italy
1959	Resumption of foreign business activities with the opening of an office in Paris

Fig. 13: Main facts of Allianz' history (from [Alli06])

Using the same source of information we can expect to find to find a similar reflection of the main facts from knowledge-extraction methods based on the management reports. The other source of information is the stock exchange.

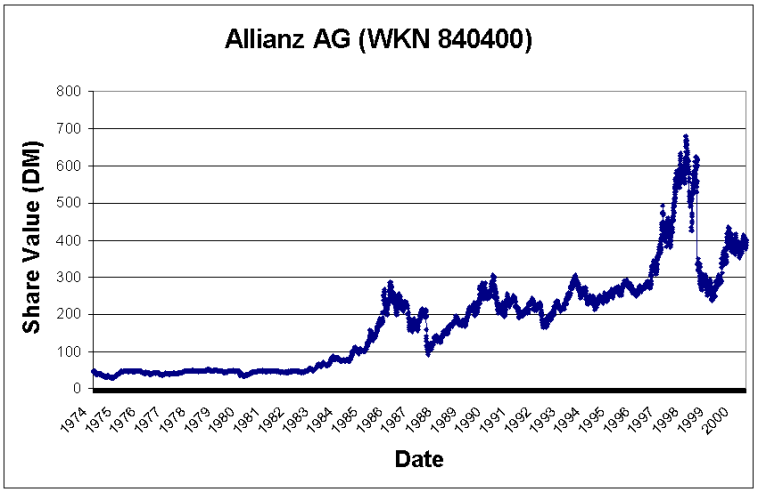


Fig. 14: Allianz' share in traded value without any clearing (source data from KKMDB)

Beginning in 1974 the shares were traded officially at the Frankfurter Börse and the values are tracked on a trading day basis. Based on the original data from Prof. Göppl[10] from the Technical University Karlsruhe, a graph was created (see Fig. 14), which documents the dynamic development of the share price beginning in the 1980s. It should be possible to follow up this development and the underlying fundamental facts when TDM is applied to the management report collection.

3.1.5 Remarks to the semantic benchmark

Due to the inhomogeneity of the introduced test sets a standardization of semantic evaluation is proposed here that allows an inter-corpus comparison.

Every concept or term can be tracked over time using a kind of counting, on the one hand, and a documentation of first and last occurrence, on the other. This information allows to describe the progress of each term and aggregated concept and will be the focus here as a basis of an expert evaluation, rather than an exact mathematical measurement.

[10] Prof. Göppl Chair of Institute for Decision Theory and Operations Research at the University of Karlsruhe, maintaining the KKMDB: „Karlsruher Kapitalmarktdatenbank"

3.2 Introduction of relevant aspects and methods for the TMF process

For the multi-disciplinary approach, introduced in domain progress extraction based on segmented corpora, several research fields do add valuable methods. The most important theories and methods developed will be introduced in the following chapters.

3.2.1 A cost function for domain knowledge extraction

Knowledge-extraction procedure costs (c_{KE}) can be interpreted as a result of an input / output production function (see formula *0*), where the input primarily are the source data acquisition costs (c_D), the costs of the extraction technology with its methods and theories (c_T) and the manual effort, e.g., editorial work and quality assurance (c_M).

[1] $$c_{KE} = c_D + c_T + c_M$$

The output is the extracted knowledge, concepts and topics that were found to be domain specific and typical for a certain time period. The quality of the extracted knowledge may vary depending on the input factors. Due to budget limitations the combination of input factors must be done according to an economical rational. The optimum is then defined as the cost minimum combination of input factors. Whereas the financial costs of knowledge-acquisition projects (from texts) are easy to determine, measuring return on investment is not easy to calculate. Ferrari [Ferr05] gave a proposal for measuring ROI in text-mining projects. They consider the specialist character of text-mining projects with intangible but present benefits. For this they differ between hard and soft ROI, where hard ROI covers classical financial aspects and soft ROI more intangible factors like customer satisfaction. The implication of the work introduced here is on the budget. As there are intangible factors present, but of worth, perhaps the expenses incurred for c_{KE} t can be higher at the optimum when the intangible factors are counterbalanced in money. An abstract comparison of various levels of c_M and their output quality in knowledge extraction is one of the main tasks of this text beginning in

Chapter 4. No exact costs are calculated for this but the assumption is made that the more sophisticated the pre-processing is, the higher the costs are.

3.2.2 Methods for data-quality evaluation

Data quality in the context of this work affects the whole processing of the data itself and of course the results of extracted knowledge. Known problems with data quality (DQ) are:

- Correctness (e.g., keying error)
- Consistency (e.g., contradictory details)
- Completeness (e.g., missing details)
- Redundancy (e.g., duplicates)
- Homogeneity (e.g., different formats for the same contents)

Before the next TDM steps can be taken, the data that may result from different sources must be processed for elimination of all DQ issues. An example of two data sets about the same person coming from different sources illustrates this in Fig. 15:

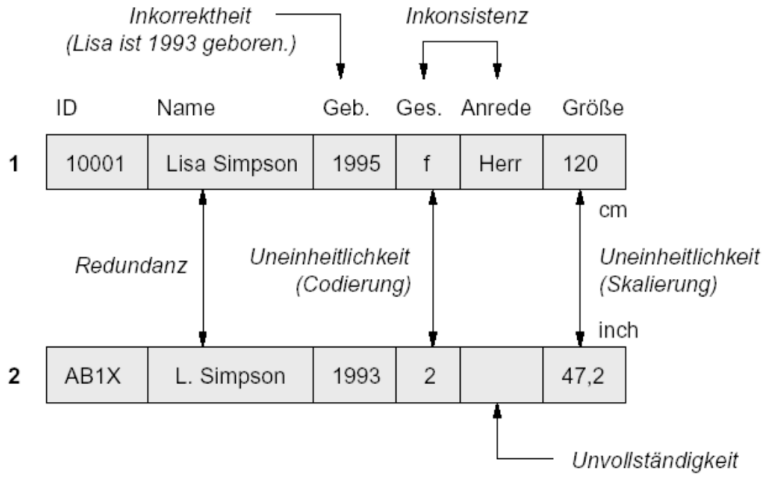


Fig. 15: Example of data-quality issues for data sets from different sources (from [Hinr02], p. 6)

In a real-world scenario of continuously occurring data that has to be processed, a continuous DQM process is to be defined. Popular buzzwords for such approaches are Total Quality Management (TQM) or Total Quality Con-

trol (TQC). TQM processes consist of a closed cycle of planning, doing, checking and acting (see Fig. 16).

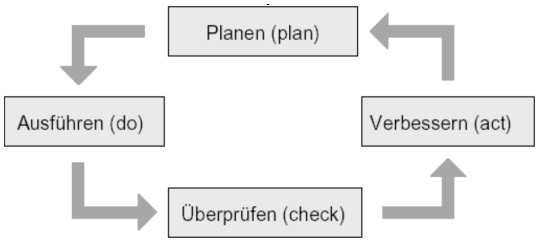


Fig. 16: Plan-Do-Check-Act Cycle (see [Tagu04], p. 390)

A common representation of source data is a basis condition for the TQM process. All TDM steps rely on pre-processed data. Any deficiency in pre-processing potential leads to distortions in knowledge extraction. The term “Common representation”, as used in this work, signifies a standard format, on the one hand, and not one-sided biased knowledge within these sources. If the source of texts is the WWW, then all markups must first be removed. Depending on the specifics of the source, it will be necessary to apply other data-cleansing steps. The specific application of certain methods – the intensity of pre-processing – will be in focus in later chapters.

3.2.3 Text data mining

Text Data Mining is a fast-growing field of interest. The main reason is the excessive growth of available textual data in companies and other non-profit organizations. E.g., Chamoni [Cham99c], pp. 355 and Küsters [Küst00], pp. 95 are introducing general approaches for applied DM that allow pattern search on databases. In this chapter an appropriate TDM process is to be defined that can be applied to data that is pre-processed on different levels of effort (and of course on different levels of resulting quality). The need for the method to be defined is therefore more a kind of universality than a high-level canon of methods. Features are the qualities or perspectives a text is analysed for or the results they are translated into. Many approaches are very general and therefore they have to focus on the task of extracting the “right” features out of the opportunities. Text-related DM algorithms are described by Dörre et al. [Dörr99], pp. 9 and [Dörr00], pp. 465. They especially introduced

general feature extraction for TDM approaches that do not consider previously available knowledge, e.g., certain analysis aims. Due to the need of setting up a relation between considered features and the whole text input they point to the need of an appropriate feature-extraction process and high processing speeds. Desjardins et al. [Desj05] introduce a genetic algorithm that is capable of knowledge accumulation in repeated extraction processes. To enhance the feature selection process the use of a pre-structured catalogue of criteria's is proposed by Doan et al. [Doan05]. Based on the definition of data mining 0 Hidalgo [Hida02], p. 2 extended this with specialities in KDD processes with text as sources:

> *[j]* "Text (Data) Mining (TDM) is the nontrivial extraction of implicit, previously unknown and potentially useful information from textual data."

Alessandro Zanasi see ([Zana05a], pp. xxvii) proposes the following definition of text (data) mining:

> *[k]* "Text Mining is an interdisciplinary field bringing together techniques from data mining, linguistics, machine learning, information retrieval, pattern recognition, statistics, databases, and visualization to address the issue of quickly extracting information from large databases."

He states that this field of research is closely driven by its applications which spread from the field of "Competitive Intelligence" and enables companies to decide strategically how to attain an advantage in market competition. Furthermore, it facilitates the understanding of people's (especially customers') behaviour and applications that enable the handling of an enormous amount of textual data in knowledge-management systems. This definition is near to 0, but it lacks the proposition that TDM has to extract "previously unknown" knowledge. Zanasi's definition addresses a more information access view,

but not really a knowledge-generating view of TDM. Therefore, I prefer the following definition:

> *[I]* "Text (Data) Mining is an application-oriented nontrivial extraction of implicit, previously unknown and potentially useful information from textual data, making use of methods from different research fields."

Several researchers, such as U. Y. Nahm from the University of Austin (Texas) [Nahm01], have proposed methods for the generation of topologies of TDM and related fields. One possible comprehensive multi-perspective view of TDM and the main terms is shown in Fig. 17.

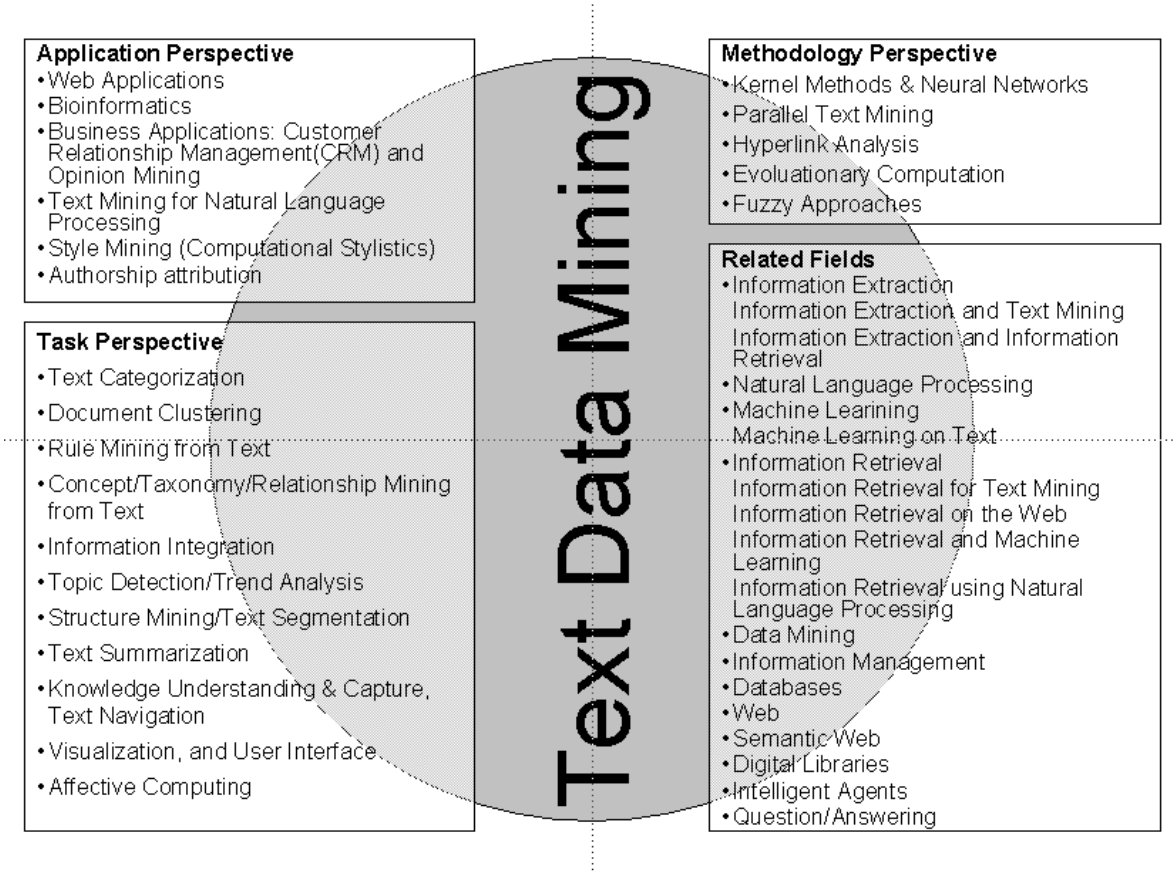


Fig. 17: A multi-perspective view of text data mining and related research fields (inspired by [Nahm01])

The *Application Perspective* addresses the real world use of text data mining which spans from genome projects in Bioinformatics to Style Mining for de-

tection of persons with similar backgrounds for terrorist network detection [Zana05a]. Other examples are introduced by Sullivan (see [Sull05], pp. 145).

The *Task Perspective* refines some of the popular data mining methods (see Fig. 3) and adds text-specific tasks like Text Summarization as well as Topic Detection and Trend Analysis and prediction (see Weiss et al. [Dame05], pp. 129). A popular task is the development of methods for text categorization, where e.g., Sebastiani ([Seba05], pp. 109) and (Mladenic [Mlad05], pp. 131) provide general work and Khordad et al. [Khor05] introduced a hybrid method, especially focused on HTML documents. Another task is classification of text documents according to key terms (see Karanikolas et al. [Kara05]). Other methods can also be used in Web mining (tracking behaviour of internet users) and prediction (see Meyer [Meye02], pp. 653).

From a *Methodological Perspective* the used techniques and approaches are focused. These approaches are not only text related by nature, but also used in several fields in data analysis.

Several related research fields that are related to the core research field of TDM surround text data mining. The distinctions regarding the assignment of each method vary from author to author. One evaluation can separate real text data mining from related approaches. Based on [b] the quality of found information has to be "…previously unknown". Hearst stands for a very strong interpretation of this definition. I agree with him and share his point of view that only "novel nuggets" are results from real data mining or text data mining, respectively (see Table 2).

Table 2: Classification of data mining and text (data) mining applications (adopted from [Hear99])

Finding Patterns		Finding Nuggets	
		Novel	Non-novel
Non-textual data	Standard data mining	AI discovery systems	Database queries
Textual data	Computational linguistics	Real TDM	Information Retrieval

"Novel" in Hearst's opinion does mean, not even the writer knows about the knowledge that is extracted from his text. An example of this would be the

discovery of an absolutely new, potentially effective treatment for a disease by exploring scientific literature. That means that text (data) mining is not information access but relies on it. Other critical work can be found at Politi ([Poli05], pp. 209) who discussed implications when applying text mining within an intelligence environment. Another general introduction to information extraction is available from Pazienza (see [Pazi05], pp. 47).

The challenge in TDM is to uncover the implicit knowledge which is hidden in the unstructured data. For this, the textual data must first be converted into an appropriate (tabular) format to permit proceeding with data mining methods. Regarding their degree of human invocation Hidalgo categorizes several classification tasks within text data mining (see Fig. 18).

	Words	Documents
Supervised learning	POS Tagging, Word Sense Disambiguation	Text Categorization, Filtering, Topic Detection and Tracking
Unsupervised learning	Latent Semantic Indexing, Automatic Thesaurus Construction, Key Phrase Extraction	Document Clustering, Topic Detection and Tracking

Fig. 18: Text Classification tasks (from [Hida02], p. 10)

Fully automatic approaches that generate results without any user interaction represent a special challenge. The problems with mining textual data are language related: This kind of data is always language specific; it contains not only semantics about a certain domain. Realizing that, methods for clustering corpora into domain-related elements, on the one hand, and language-related, on the other, are needed here. This task is not straightforward due to the polysemic character of many terms.

3.2.3.1 Clustering and naming

Most of the known methodologies need a domain expert either at the stage of pre-defining categories and cluster names or when selecting topics or cluster names from a semi-automatic system. Here an approach which automatically

extracts possible cluster names out of the given data set can be helpful. This can be done by observing general time-dependent measures of the given corpus and focusing on such terms that appear at periods which have been identified as worthy of analysis.

In repeated scenarios there is the chance to accumulate knowledge during several iteration steps. Borgelt et al. [Borg04] using weighted key figures for the extraction of most important keywords. With such methods it is possible to consider previous knowledge from prior DM steps or knowledge that was brought in from external sources or the user himself.

One major focus of research is clustering[11], which simply follows the aim described briefly here: Given an unknown set of different text collections the task of grouping those according to inherent qualities in order to produce clusters of different qualities that contain similar documents. This weighting of keywords is one methodological approach (see Frigui et al. [Frig04], pp. 47). A more cross-method approach is introduced by Salazar et al. [Sala05]. Their method derives association rules from clustering processes. A task-oriented clustering was introduced by Santos et al. [Sant05] who use the extracted knowledge for database marketing applications. Approaches spreading to social- and society-focussed research like Vafopoulos et al. [Vafo05] are ongoing, in which they offer a method of distributed “HyperClustering” for overcoming the digital divide.

The clustering in the approach introduced here is based on quantitative and statistical qualities of the terms found in the test sets. The application of TRQ measure allows clustering terms according to their importance within a certain text collection. The persistence quality of each certain term allows clustering from this perspective. Additionally, a semantic grouping was applied using domain-specific taxonomies. The basic ideas behind that are introduced in the following chapter.

[11] For an introduction see Mandreoli et al. [Mand05].

3.2.3.2 Progress extraction and topic evolution

In several research fields the metaphor of “Waves” is used to describe progress paths of social or business developments, research topics and product lifecycles. Classical examples are the “Kondratieff-Waves”, which are used for the description of the social development within the last 200 years, beginning with the industrial revolution. Nefiodow ([Nefi96], pp. 126) defined a 5th Kondratieff wave, which is no longer based on the processing of goods, but on services, consulting and knowledge. Such a kind of paradigm change is the subject of many research activities. Not only societies but also research fields may be seen under this “wave” paradigm. In this chapter I focus on research that has in common the aim of extraction of knowledge and their semantic change over longer time periods, but the differences from my metadata-based approach will be worked out. The focus here is more on the method than on the application as a whole, which widens the view regarding non-commercial research activities that have not produced a TDM Tool yet. Current approaches for trend-detection or trend-tracking systems can be divided into two main classes: Fully or semi-automatic, depending on the needed involvement of a domain expert in the detection or tracking process. A detailed analysis and evaluation of commercial applications in trend detection is provided in [Kont04], pp. 185. One popular way to follow up developments in technology is patent mining. Larreina et al. [Larr05] give an overview on state-of-the-art tools for patent analysis, which allow monitoring developments and delivering expert information. They introduce main bibliometric analysis methods based on measures of co-existence of words and their statistical foundation, the intensity of appearance and building technology maps from the results of analysis. This approach can be extended by quantitative measures to enable decision making regarding the importance of a certain concept. The focus in this work here is more on the definition of appropriate thresholds for significant appearance of concepts than on the analysis of their co-existence.

Other approaches focus on the question of how documents and trends can be related in a way that it is possible to assign new documents to certain progress paths or trends. Lavrenko et al. ([Lavr00]) introduced EAnalyst, an

implementation of a general architecture for the task of associating news stories with trends. The idea is to predict trends on the basis of news articles actually published. Different to the focus of their work (assigning documents to build models via likelihood criteria), the work introduced here breaks the structure of single documents and only uses quantitative measures of tokens. Web queries as another source for trend and behaviour detection and tailored methods, Wang et al. ([Wang04], pp. 176) are introduced as an alternative approach.

Lebeth et al. ([Lebe05b], pp. 271) do have a comparable aim to the approaches that are introduced in this work. But the focus is different. They are developing methods and prototypical implementations of a text-mining-based knowledge-management system in the banking sector. The idea is more to organize knowledge in huge archives with a common access to that than to generate new knowledge based on these sources.

3.2.3.3 Topic detection and tracking

In this chapter two research directions that deal with the task of extracting new topics or events will be introduced: Topic Detection and Tracking. Both have in common detection with appropriate methods from texts. Perhaps a “topic” can be pre-defined, e.g., “terrorism” but the assigned stories do not have to contain the term “terrorism” (but, e.g., “bombing”). New documents are analysed whether they belong to one of several already detected topics or to a previously not known new topic. The difference between “topic” and “event” is that an event is an instance of a topic. In this meaning the topic is the top-level label. The challenges here lie in the recognition of whether an event belongs to a known topic or if this event may initiate a new class of events (it is a new topic in this case). A learning capability is always needed for both topic detection and event detection.

In the mid-1990s the research activities get pushed by carrying out the “Topic Detection and Tracking” (TDT) task by a few research organizations and the U.S. government. The subject of this research is event-based organization of broadcast news ([Alla02a], pp. 1), ([Alla02b], pp. 197). The main research

task was divided into the following sub-tasks: Story Segmentation, First Story Detection, Cluster Detection, Story or Topic Tracking and Story Link Detection. Most of the methods used are bottom-up approaches that analysed text corpora word-by-word or sentence-by-sentence and used clustering and tagging techniques [Spil02]. Other methods are more statistically based and work with Features, e.g., corpus-wide measures or Vector Space Models which represent sentences or whole stories. Applications based on these methods are realized, e.g., for Automatic News Summarizing, Document Clustering and Patent Mining.

TDT belongs to the class of unsupervised learning. Two fields of application exist: The retrospective detection from a historical collection of texts and the detection from news streams in a kind of real-time scenario.

The used methods are, e.g., clustering techniques based on vector space models of the corpora, citing [Yang99], pp. 3 and [Yang02], pp. 85 as examples out of many and developed clustering techniques based on TF-IDF measures combined with similarity measures based on cosine and k-nearest neighbour clustering algorithms. Comparable approaches for search in text collections can be found in [Koba04], pp. 103. So they deal with the deficiencies that vectors of documents are not really orthogonal and, therefore, cosine similarity measures are not fully appropriate. Senellart [Sene04], pp. 25 criticized that vector axes of a document collection are seldom fully orthogonal and the documents tend to have something in common and therefore they are not fully independent.

Montes-y-Gómez [Mont99] introduced a method to discover information that uses a classical statistical model based on distribution analysis, average calculus and standard-deviation computation. The goal is the extraction of social topics out of news articles.

Other approaches that use neural network, e.g., [Raja01] to fulfil similar tasks, are also available. One can subsume that the application of certain DM algorithms is not task specific, but more or less theoretically well founded by each particular researcher.

The treatment of individual documents out of a set of related documents is another aspect. One class of techniques makes use only of internal semantics, e.g., the occurring concepts. That is especially true for TDT applications. Another class of methods is based on measures of explicitly given relations such as links or citations of external articles.

There are several frameworks available, created for following up on TDT tasks, e.g., HotMiner ([Cast04], pp. 124) a tool for the semi-automatic extraction of pre-definable topics. Other implementations on the basis of probabilistic approaches Leek et al ([Leek02], pp. 67) are introduced while Yamron ([Yamr02], p. 115) et al. focuses on the creation of statistical models of contents. Contrary to trying to discover new topics or events from given text collections, other research is focused on tracking changes compared on the basis of textual sources released before and after a given event. One example is that of Rehbein ([Rehb04], p. 85) regarding the word use before and after 9/11[12].

These examples in research (briefly introduced here) show several specialised research directions under the headline “topic detection and tracking”. They are united by the fact that the focus is on the contents with taking the quality of sources as given and quite theoretically “optimal”. But this is not the case with real-world scenarios, especially when a mixture of documents is used as sources that are of different pre-processing quality. With this ex-ante assumption of an “optimal” source all previously mentioned approaches differ from the approach introduced in this work which proposes a set of statistical measures for qualitative classifications of given text collections.

3.2.3.4 Literature mining

Special attention of research in literature mining is given to approaches that allow extracting and relating facts automatically from scientific publications or detecting events in historical sources. This field is called literature mining. De Bruijn et al. [Brui02] introduce a general process in literature mining and re-

[12] 9/11 is a common abbreviation for the terrorist attacks against the World Trade Center on 11th September 2001.

search from several fields. Their focus is especially on the extraction of medical facts from online publications, but the introduced process itself is of a general quality.

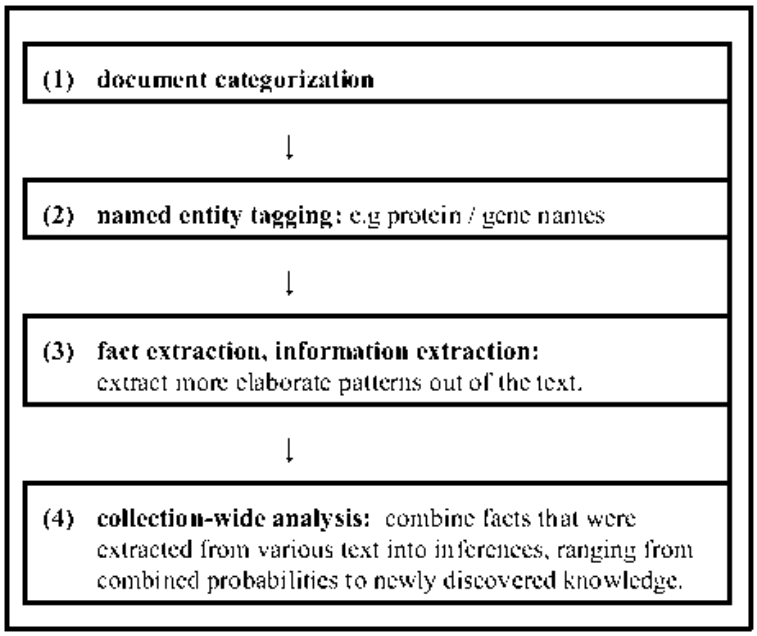


Fig. 19: Process in literature mining (taken from [Brui03], p. 556)

The process in literature mining (see Fig. 19) starts with a document-categorizing step. Further processing follows the aim of enriching entities with facts within certain texts firstly, later within the whole collection.

Within the Perseus Project at Tufts University (see [Smit02]) an approach based on rank lists of collocations was developed for detecting and browsing events in unstructured text. Terms are ranked by chi-squared values. The results allowed a line of events and historical places to be followed on a geographical map which showed where events in the U.S. Civil War took place. Whereas in the Perseus Project the driving questions are “What happened? Where? And When?” the dimensionality within my approach is lower, because the collocation with the geographical location is not in focus. Here the “What? And When?” are assigned by the given information of issue time for the time-segmented corpora.

3.2.3.5 Complexity reduction

One of the main issues with textual data is the high dimensionality. Howland and Park propose a vector-space-based approach for dimension reduction ([Howl04], p. 4). This approach focuses on a similar goal – the improvement of efficiency in textual data mining – but makes use of different methods. The processing is mathematical reduction of the rank of the vector space matrix. In contrast to my semantic segmenting approach Howland et al. apply their dimension reduction in the phase of data mining within the DMP. The advantage of their proposal is the flexibility in application with different research objectives. The gaps are approximation losses of semantics in the data and the non-optimal usage of the semantics within the data because the method is on a pure mathematical level and also potentially non-task-specific data is processed.

3.2.3.6 Semantic evolution

E.g., Baron et al. [Baro03] discussed several aspects of pattern evolution. In general two main types of changes of the contents of a sample are to be observed, i.e. the connection which is described by the sample and changes of the measures the sample is described by. For tracking changes the data must have a time dimension. An appropriate selection process for time granularity is necessary to guarantee a complete extraction of rules from the test data: Too-large intervals in partitioning lead to incomplete rules. Several methods are available to determine appropriate periodicity, e.g., rule based or formal. Baron et al. introduced a pattern monitor (PAM), which uses (non-directed) mining procedures only in the first step, followed by applications of (directed) SQL statements in the following steps.

Classification problems with concepts extracted out of large text collections may also occur. On the whole, one term can be assigned to more than one category. The membership of each particular term is not clear to define, e.g., multiple memberships may describe this situation more properly than the assignment of only one set. Fuzzy Logics may help in these kinds of problems. Fuzzy theory is an extension of the set theory introduced by Lotfi Zadeh in

1965 as a new theoretical paradigm in social research. The main idea is the introduction of a “linguistic variable” that enables an easier projection of real-world scenarios into a set representation. Fuzzy sets are not randomly based, but on vague definitions of sets. The greatest successes that fuzzy logics had in the 1980s and 1990s were found in application-oriented use cases (see introductions in [Kahl93], [Klir95]).

Since the beginning of the new century a kind of renaissance of Zadeh’s ideas [Zade99, Zade02] and various extensions have led to several new approaches, especially in the field of TDM.

Alemáo et al. applied fuzzy logics to neural nets for financial forecasting [Alle05]. Other approaches to allow to model smooth semantic changes, e.g., clustering of terms or concepts to more than one cluster. Doing so also allows tracking different progress paths of meaning in parallel (see example in Table 3).

Table 3: Fuzzy cluster assignment of terms

Term	Time stamp	Measure (TRQ)	Cluster Computer networks	Cluster Electronic Communications	Cluster Mobile Communications
Mailbox	1985	12	0.9	0.1	0
Mailbox	1995	16	0.3	0.6	0.1
Mailbox	2005	13	0	0.4	0.6
T_2	…	…	…	…	…
T_n	…	…	…	…	…

Kwiatkowskal et al. [Kwia05] presented an evaluation of clinical prediction rules using a convergence of knowledge-driven and data-driven methods, a kind of hybrid approach between specialist directed and algorithm driven, as they call it, a “semio-fuzzy approach”. These examples show the variety of applications that are enabled by the simple “fuzzy” idea. Straccia ([Stra05], p. 1) worked on the problem that description logics (e.g., OWL) that are usually used “…becomes less suitable in domains in which the concepts to be represented have not a precise definition.” One example is the flowers domain, where the description of objects often relies on adjectives for certain qualities like colour of trees or thickness of the handle. Within the current research the fuzzy quality of concepts is considered in the interpretation step of results,

not in processing. Here an evaluation with expert knowledge is applied. For more detailed information on handling fuzzy qualities refer to the original literature.

3.2.3.7 Human-driven methods

If the share of computerized work in analysis is minor, the approach is called “manual” here. One influencing activity is introduced further on.

The honoured German computer scientist Prof. Peter Mertens analysed the development path of the scientific discipline of business informatics in the mid-1990s ([Mert95], p. 25). He proposed that a scientific discipline might develop along one of three theoretical progress paths.

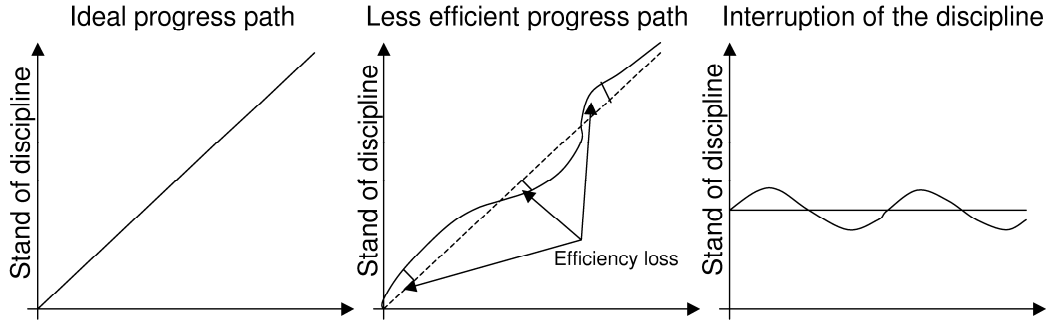


Fig. 20: Possible progress courses of a discipline (translated from [Mert95], p. 25)

Mertens stated that meandering developments with efficiency losses, like that shown in the second graph, cannot be prevented and may lead to important experiences within the discipline. The goal of his analysis was to find out how and along which progress path the discipline of business informatics develops on an aggregated level. At a granularity below he wants to find out which topics belong to which path. The methods used were manual: A counting of articles of the German weekly publication “COMPUTERWOCHE” according to predefined categories, which were dynamically adapted to new topics that occurred in the publication over the time period of 20 years. This research activity is very close to my dissertation from the point of view of motivation. He also wanted to learn more about a certain domain by the use of textual data. Mertens himself criticized the method he used, especially that only a manual counting of articles was applied. Neither did he use any data mining

method, nor were all available articles completely considered, but only the headings and only selected contents of the articles themselves. The differences to my work are the absence of any computer-supported analysis, the incompleteness and the latent individual biasing during the analysis process.

3.2.4 Computational linguistics

The research field that deals with specifics of languages with the support of computers is Computational Linguistics and is briefly introduced within this chapter, especially regarding differences to text-data-mining methods.

Approaches in linguistics are chiefly based on methods that work on each text file itself for clustering, tagging or classifying purposes [Lend98, p.130], [Moen00]. The main objective is to support information retrieval or later querying against a mass of these text files that were processed the same way.

Another popular task is authorship attribution and stylistic analysis. The aims here are to assign texts to authors or genres. This can be done, e.g., by the use of statistical models [Baay93], corpus measures [Baay02] or without lexical measures (see [Stam01] for an introduction).

The usual methods in linguistic processing are truncation, stemming, lemmatization and tagging. These main procedures will be introduced here only on an abstract level, only to establish a general understanding for later deciding which methods will be used for the approach within this work.

- Truncation means to cut off a certain number of signs, e.g.: {Museum, Museen, Mus, Muster} → mus
- Stemming acts according a strict algorithmic logic for the detection of word stems, e.g.: {Museum, Museen} → muse
- Lemmatization is the reduction of a word form to its base form (lemma), e.g.: {Museum, Museen} → Museum
- Morphological analysis uses more defining attributes for characterization, e.g., (gend=neutr, nbr=plur, case=u)

Further on, the focus of pre-processed text collections may be the machine translation of a text. This method for word sense disambiguation is developed, e.g., by Miangah et al. [Mian05] by applying a large target language corpus.

Here machine translation is only used for parts of the Al1k corpus. However, the focus is not this translation process, but the target corpus that then exists in the German target language. The translation itself is taken as a “black box”.

Different kinds of knowledge domains must be handled in different ways. For example, technical domains, e.g., business informatics, have special qualities, which make it necessary to configure the methods of research to the needs of the research aim. If the basis for research is a huge collection of documents from a more technical domain, such corpora must be handled differently from literary corpora. For example, product names, programming languages and other proper names must be kept during all analysis steps. Pure linguistic approaches therefore are not applicable without enhancements. As the aim is to extract new knowledge (also new concepts and terms) an application of algorithms to the terms found was rejected. Terms will be taken “as is” in following processes. The filtering of garbage data will be done using an automatic approach based on applied taxonomies and statistical threshold measures, not by the application of a predefined stop-word list. Most of the used procedures do not consider the time dimension and do not establish a time perspective between several dates of a term occurring.

In contrast to these methods, introduced here shortly, the basic idea of the segmenting approach for tracking progress paths is the assignment of each single source text to a time-based corpus segment.

3.2.4.1 Text (data) mining and computational linguistics

Contrary to data mining with the aim of generating new knowledge, the focus of computational linguistics is the explicit materialized language[13] and their usage. The semantics are interesting as long as it belongs to the semantic concepts within the texts. Generated semantics itself (generated by applied algorithms) is not the focus within this field. To support the process of Automatic Detection, Classification and Visualization of Trends, the main task is to extract the semantic concepts out of textual data and separate them from language specifics.

Computational linguistics deals primarily with the development of computer-based methods which help to answer linguistic research questions and Natural Language Processing (applications, tools etc.)[14]. The object of interest is not a special knowledge domain (except that of computational linguistics, of course).

The experience object within CL is a “corpus” that consists of tokens (elements of the corpus). A linguistic corpus is tagged (enhanced with linguistic information). For tagging of corpora different techniques and methods are used. The Text Encoding Initiative (TEI)[15] proposes one example out of several approaches for tagging. The TEI standard is based on SGML with several derivates defined in SGML family languages, e.g., XML. The automatic analysis of electronic texts has a long developmental history which dates from the middle of the last century. Modern research approaches use statistical methods for quantitative analyses of text corpora, e.g., for semantic simi-

[13] whether in written or spoken form

[14] see [Smit91] for an overview

[15] http://www.tei-c.org/

larity checks of document sets. The linguistic methods therefore can be separated into two main directions of research activity:

- The predominant quantitative focus that uses corpus measures of text corpora, e.g., term or word frequency for evaluation and comparison of different text sources, see [Atte71] for examples.
- The more qualitative focus that concentrates on word use, structure and semantics, see [Lend98, p. 106] for examples.

It can be observed that (comparable to DM) most of the methods for text data mining work bottom up, starting at the smallest unit of textual data, a word or n-gram (only a few letters) and generalizing the pattern which were found. For the tracking of trends over time in technical domain corpora these known bottom-up procedures have limitations:

- A lack in performance may occur when a real large corpus is analysed at a time not in direct access (cannot be stored in memory).
- The patterns found are based on generalized results of multi-parametric algorithms which mean that a biased result is to be expected due to the multiplication of error terms.
- Language-elements-orientated approaches are not appropriate, because technical information[16] is not considered or is excluded within these analysis processes.
- The generation of action recommendations is not transparent to “normal” users.

Corpus linguistics is an area of linguistics which creates or tries to prove theories about language using examples from text corpora. These corpora can be small for specialised research questions and large for the search for more general and statistically significant rules. This field faces a text as a whole with the focus of the empirical foundation of linguistic theories. Corpus linguistics is an inductive/empirical method for the profit of knowledge about the language: One puts forward a theory after the observation of as many

[16] e.g., abbreviations for norms and techniques

individual examples as possible. That is in direct competition to the deductive method and which earlier was practically the only valid method in linguistics and is derived from the philosophical tradition of linguistics to date: The scientist wonders how language is built up and tries to find examples of his consideration in more languages. Corpus linguists developed several measures which are used as so-called “markers” for diverse stylistic tests, e.g., in author attribution applications

The “physical” research object differs between the main qualities of CL and TDM, especially if the task is trend extraction from large text collections. The following challenges in TDM are to be especially considered:

- Incompleteness of text collections
  A corpus is finite. All a writer wanted to say must be written explicitly in this text. Otherwise it cannot be considered during the analysis process. In real-world research scenarios perfectly valid corpora are rare. Collections of texts that are related to a certain domain can only reflect semantic parts of this domain. In TDM, especially in the current approach, discussed later on in this text, research objects are large text collections, which can be seen as a “window” of a number of terms out of all texts that may also be related to the knowledge domain that is to be observed.
- Different authors
  Different authors have their own individual style. Different authors have their own opinions about certain things. Only the domain they are writing about unites them. Using large text collections of different authors as sources of knowledge extraction should therefore lead to a smoothing effect regarding individual and/or subjective reflection of reality.
- The CL terminology does not fit
  Classifying and counting terms is very common and important in CL. Research questions may focus on rare items, e.g., hapax-legomena and dislegomena (these are terms that appear once or twice within a corpus), or frequent items. Due to the “window” or subset quality of the texts that are processed it is not possible to exactly classify if a term is a hapax-legomena or dislegomena. This can only be done in relation to a certain

corpus, but that may be incomplete. Conversely it is crucial in trend extraction and TDT not to select a corpus subset that is too small to detect emerging trends, like “XML” in the mid-1990s.

Due to these characteristics of original CL methods it is necessary to apply an extension in the methods for keeping terms unchanged during the extraction process of domain knowledge progress. This approach is introduced further on.

3.2.5 Knowledge representation

Knowledge can be represented in different levels of formal degree. The most common levels are (see Fig. 21):

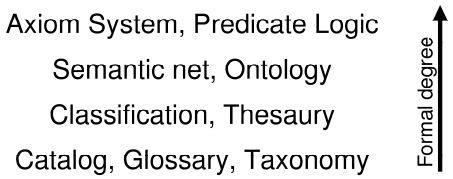


Fig. 21: Knowledge Representation Approaches and their formal degree

An additional aspect is the ability to recognize knowledge. One subset of whole knowledge is the knowledge that is documented, made explicit in some kind of representation (see Fig. 21). The representation can be

- Humanly readable
- Unstructured data (audio, video, plain text)
- Machine readable
- Structured data (Ontology’s, Database Format, XML, HTML)

Kinds of materialized occurrences can be lexicons (see [Hirs04]).

Although data is machine readable, the challenge of combining different sources remains. Witt et al. are working on unification of XML documents [Witt05], which make use of concurrent markup. A convergent representation of the semantics contained within the source data is the basis of reliable results of further DM steps. The task is to bring together extracted knowledge

from various extraction processes. There are several approaches available, from Noy [Noy04] and Doan [Doan04].

Another subset of the theoretical concept of “whole or infinite knowledge” is implicit knowledge. Some knowledge is manifested in processes, “between the lines” or in the minds of people. In the field of knowledge management it is one of the greatest challenges to uncover such individual or hidden knowledge. Approaches known in the field of knowledge management to turn implicit knowledge explicit and make it available for companies (see [Stei00]).

Other examples of such knowledge are, e.g., production processes, which have a very complex production function. Very experienced process engineers may control such processes intuitively, sometimes with success even if they use different parameters from person to person ([Otte04], p. 81).

Originally taken from the philosophical field, the term “Ontology” in DM describes concepts and their relations within a certain domain. The community that is working in this field created and refined several definitions, especially in the last ten years, due to the extensive usage of logical representations in various knowledge and semantic web applications. Gruber ([Grub93], p. 2) defines ontology as “an explicit specification of a conceptualisation”.

A more suitable definition is given by Studer et al. [Stud98] that describes best what ontology means in terms of use in this work:

> *[m]* “Ontology is an explicit, formal specification of a shared conceptualisation of a domain of interest.”

A formal description of ontology is given by Cimiano et al. in [Cimi03]. For a comprehensive introduction to ontology construction I refer to the original sources, e.g., [Brach04], [Gome04b] or [Abec04]. Ontology can be constructed for their semantic concepts modelling knowledge domains by the use of directed graphs, which can show relations between elements of domain corpora. Ontologies can be used in several application scenarios. For use in WWW environments a specific ontology web language (OWL) was

developed for a standardized interchange of ontology information[17]. Based on such concepts deeper analyses, e.g., the use of classical data-mining techniques is possible in later stages.

As the hierarchical element of ontologies – taxonomies – describes relations between the concepts of ontology e.g., “belongs to” or “consists of”. Vallet et al. [Vall05] define taxonomy as “the root for class hierarchies that are merely used as classification schemes, and are never instantiated.” In this work the preferred definition is:

> *[n]* “Taxonomy is a root concept that allows the classification of terms that belong to a general concept within a domain.”

Remembering Fig. 21, taxonomies represent a very simple and non-restrictive representation of knowledge. Although this representation is preferred within the current approach, the three aspects *construction*, *maintenance* and *applications* of knowledge representation will be briefly introduced here.

For the *construction* of logical representations of knowledge several methodological approaches are available that support the process of source data structuring using ontologies. To do this, several methods are under development, e.g., the resource description framework (RDF) [McBr04] with a special language, the semantic web rule language (SWRL) ([Pan05], pp. 4) for web sources or the On-To-Knowledge (OTKM) methodology for knowledge representation [Sure04].

Most of the domains do change over time. There are methods available that consider changes in knowledge representation by the use of learning algorithms for the *maintenance* ([Paaß04], [Maed04], and [Bras04]).

A wide area of *applications* is available from general knowledge-retrieval solutions (e.g., [Eklu04]) such that they are specialised in certain domains, e.g.,

[17] See [Anto04b] for a detailed description.

medicine [Hahn04]. Others focus on access to knowledge via portal-based solutions [Obe04]. An approach for an ontology-based platform for semantic interoperability is proposed by [Missi04].

The construction of the taxonomies used within this analysis is introduced in Chapter 3.6.1.

3.3 Pre-processing

The time spent in DM activities is approximately 20% on task definition and approximately 60% on the choice of the relevant data sources in usual TDM processes. Approximately 10% is spent on re-processing, interpretation, evaluation and application of the data-mining results and also only 10% on the application of data-mining methods ([Cabe98], p. 43 and similar orders of magnitude at [Küpp99], p. 117).

The five main steps in the data-mining process are not considered with the same level of attention to that relation found for the single process steps in research activities of the data-mining community.

Table 3: Returned Google results for the data-mining process steps

Data-Mining Process Step[18] (equal to the Google search term)[19]	**Returned search results**	**Proportional relation (to the term "Data Mining")**
Data Mining selection	6,800,000	13,03%
Data Mining pre-processing	*288,000*	*0,55%*
Data Mining transformation	2,460,000	4,71%
Data Mining	52,200,000	100,00%
Data Mining interpretation	3,010,000	5,77%
Data Mining evaluation	8,940,000	17,13%

A retrieval of the related terms that belong to the five steps and the general term "Data Mining" brought remarkable results. Surprisingly the "Selection" and "Pre-processing" steps, which together take up about 60% of the time

[18] Due to better comparability, the step "interpretation and evaluation" was divided into two separate search requests.

[19] The Google search was processed on Oct. 22/2005. Parts of terms were concatenated by AND operators, if the term consisted of more than one term.

spent in data-mining projects, are considered much less (especially the “Pre-processing”) than all other steps within the process (see Table 3). This “Googleling” is indeed not very robust regarding a scientific foundation but it gives a raw orientation on the trend the researchers prefer to invest their limited research budget. Experience from conferences and discussions with other data-mining researchers (together with the observation of published papers) lead to the conclusion that the focus in the data-mining community is on the development and improvement of core data-mining methods and not on the improvement of the data-mining process as a whole. Practitioners agree with the thesis that good data-mining results depend more on an appropriate data selection and a high-quality preparation of the data than on the data-mining method itself ([Otte04], p. 244). The realization of this fact initially triggered my intention in my research to focus on the first steps of the data-mining process while also considering the whole process and the result that is presented to the user. Lowering the expenses of pre-processing would significantly increase the efficiency of TDM.

Pre-processing is usually mainly seen as a more mechanical procedure of data cleansing in a Data Quality Management (DQM) process. The expected result of this step is always a set of documents in a standardized format. The aim of this work is also to extend this pre-processing step towards a task-specific step within the whole data-mining process.

The other aspect of this dissertation is the textual nature of the processed data. Whereas data that is represented in databases is easy to compute but difficult to understand by humans, texts are easy to understand but difficult to compute for retrieval, shared creation and use as well as semantic linking.

While text (data) mining is a relatively young sub-discipline, the above-mentioned phenomenon of a non-adequate representation of the preparing DMP steps is more to be found than in classical DM. As the main adoption, the textual data is converted into a vector-space model for further processing that allows applying standard data-mining procedures on the resulting tabular structures.

In contrast to the usually used bottom-up approaches the top-down approach presented here allows shortening the amount of data that is to be analysed by the DM algorithms by analysing the whole data set on a meta level first, and applying an initial filtering at the first step. It was empirically observed that the amount of data that was left for later application of DM methods was reduced by up to 80%. The left-over data is cleaned from noisy data that would otherwise bias the results of applied DM procedures on the data if the filtered data would have remained in the analysed data set. It is to be expected that the amount of computerized data is not only reduced but also the quality of the results at the end of the DM process is higher with the proposed pre-filtering approach. Furthermore, the estimation of corpus measures allows predicting future developments at domain level.

3.3.1 The method used here

The sources “Computerwoche” and “Allianz” have different kinds of origin. Whereas the “Computerwoche” originally came in electronic format from the WWW, the Allianz Management Reports are hard copy paper. The TDM method introduced in Chapter 3.5 works with terms stored in a database. The following data processing steps were applied on the data:

1. Transformation into computer-readable format
2. Extraction of domain-related data
3. Elimination of duplicates

The two sources used here need differentiated pre-processing, e.g., the CW source, of course does, not need step 1.). The applied pre-processing for both sources is explained in detail in the following chapters.

3.3.2 Pre-processing of CW

After the downloading procedures, the next step was to clear each HTML-format issued article from the media specifics and non-relevant information to

prepare further steps (see Fig. 22 for a sample HTML page with the pure article text marked by a square).

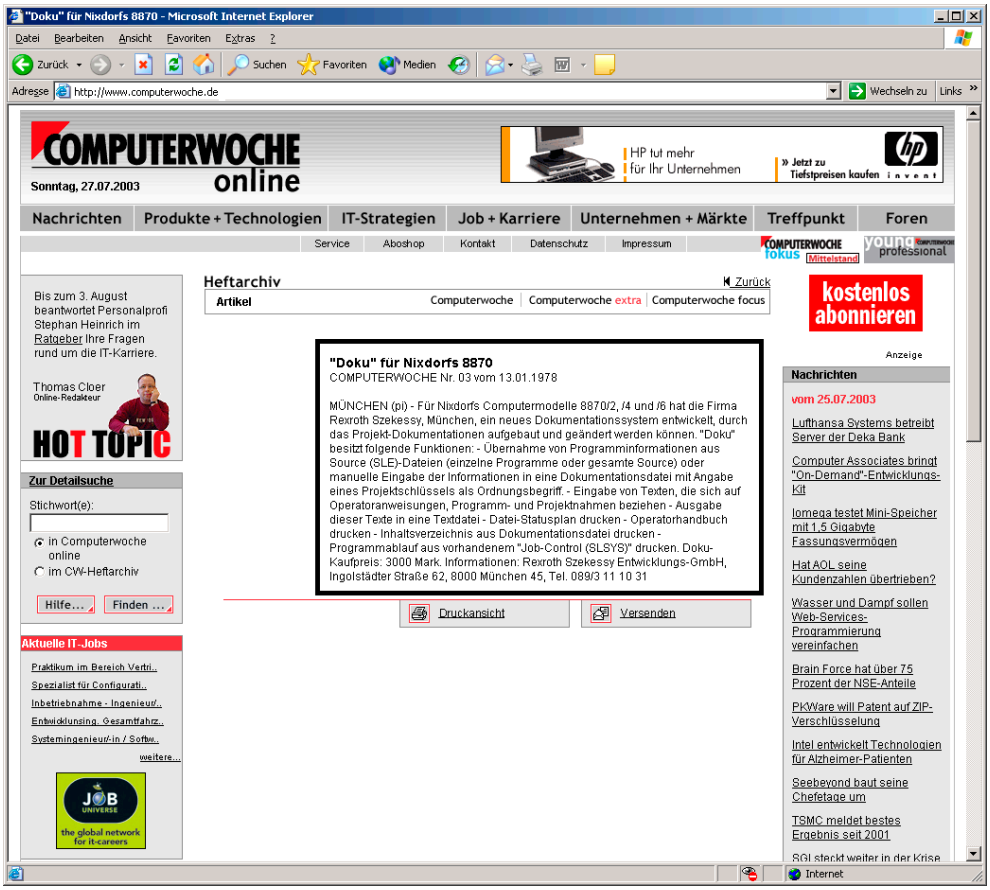


Fig. 22: Source document format with added square that marks the area of interest

The usual non-editorial information portion with web pages (e.g., advertisement, article and/or hyperlinks to other articles, which do not belong to archives of the CW) was eliminated, so that only in each case the article text was near drawn for evaluation. A parser, which eliminated all content that surrounded the article text, was used to clean the data. What resulted were files coded in simple HTML as shown in Fig. 23:

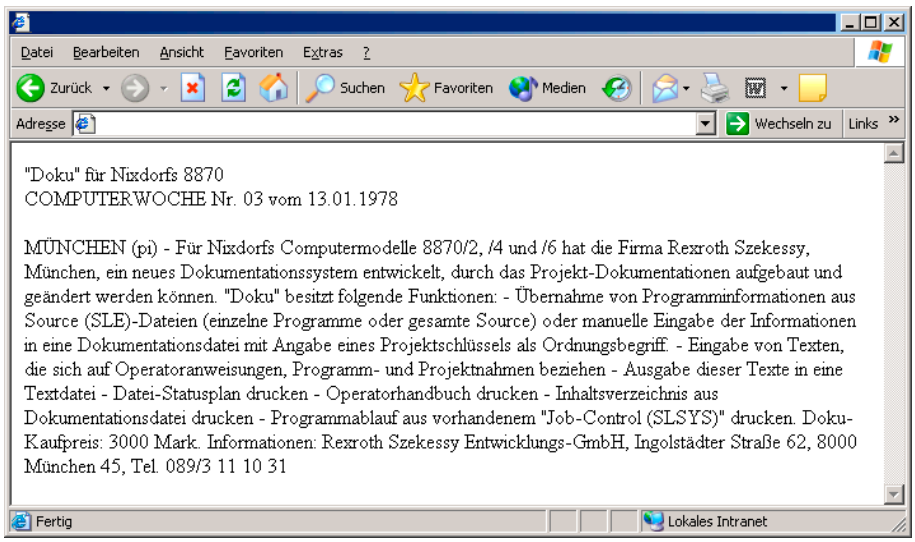


Fig. 23: "Cleaned" HTML page with remaining target data C_T

The extraction was carried out by the use of parsers written in Perl that allowed the use of complex statements (regular expressions). The main structure of each single CW article file was equal over all files. Therefore, it was possible to explicitly define the definition of semi-tags for the start and the end of the editorial content of the articles. From the resulting HTML file ASCII files were generated as a basis for storing the terms in a relational database. During this stage the original structure of the single article files was transformed into a vertical structure of terms enriched with aggregated counts of occurrence within each single article and issue date (this will be the basis for later aggregation to yearly segments, see Chapter 3.5).

In the data-mining process (see Fig. 4) and in the common approaches in the field of TDM, the pre-processing step is a necessary conversion step as preparation of the data before further steps can be properly applied. The result is a “bag of words” in a standardized format. This is not precise enough in my opinion because in the pre-processing step the internal semantic that a text collection has must be considered. Curia et al. [Curi05] introduced a sophisticated tool OLEX that supports the extraction of linguistic, syntactic and structurally relevant features and annotates them automatically. In contrast to such linguistic-orientated tools an approach is needed here that focuses the not-annotated terms, but considers their persistence qualities over time. An extended “pre-filtering” that overcomes this deficiency and allows the extraction of progress in domains is introduced in Chapter 3.5.1.

3.3.3 Pre-processing of AI1k

All available printed management reports of Allianz from 1962 to 2000 were transformed into an electronic representation using a document scanner and OCR[20] software. After this an automatic spell check was applied to the data to correct falsely recognized terms.

3.4 Conversion into a standard format

Text source files conversion to a standard format (e.g., XML), which is then used throughout all subsequent steps, must be done to eliminate all format-specific differences between all single texts. Graham ([Grah00], pp. 12) describes an example in detail of pre-processing for newsgroups' mining. He also uses internal structures, e.g., headers, bodies and others for the conversion to a standard format.

No single standard format is proposed here but the assumption is made that one format is declared to be standard and all sources are converted to this standard. That may be an XML file, structured according to a specific Document Type Definition (DTD). In this work simple ASCII was declared as standard. After the extraction of the contents from the original input files, both sources were converted into text files in ASCII format. After this, a domain-related text collection exists that is ready for further processing.

3.5 Pre-Filtering and corpus measure pattern recognition

KDD approaches (see [Fayy96]) have the aim of extracting new, "potentially interesting patterns" from collections of data. The meaning of "new" depends on the aim of the research and may have implications on the methods that are applied. E.g., if the aim of research is to track progress in domains, the extraction of two clusters – relatively persistent concepts and relatively volatile concepts – is not really new, because it is inherent in the research aim. The aim of extracting constant and volatile concepts must therefore be con-

sidered before the application of DM methods. Another role in this context is the previous knowledge the knowledge worker has about the certain domain. The significance of pattern found in the DM process depends highly on the ex-ante domain knowledge of the knowledge worker who is conducting the analysis. Pohle [Pohl04] in his work focuses on the individual significance of the pattern found and proposes methods for integrating domain knowledge into post-processing. In contrast to this work, the individual knowledge of a certain knowledge worker is not addressed here, rather the aim of completely extracting concepts regarding their persistence quality within the domain. Individual aspects may be considered based on the pattern found with the proposed methods here, but are not focused on within this work. I propose that by considering the internal semantic of the data within these first steps, better results can be achieved in the following TDM. By that, the quality of the pre-processing is not meant here – this must be done carefully, precisely and by applying the methods with high quality – but if the task of TDM allows making use of a semantic pre-processing then this must be done before applying DM methods.

Most of the researchers today make use of a “generalizing” paradigm: converting every source data into a common DM format (the tabular or matrix one) and applying a canon of special or standard DM methods to this standardized data. Drastically stated, that applies to the status quo in most DM approaches. If there is this *general* paradigm, a *task-specific* paradigm should be possible to construct. Basic principles of this idea are introduced in the following chapters.

3.5.1 Task-specific segmentation of text collections

The result of the previous pre-processing is a “bag of words” in a standardized format. Until this point the approach presented here does not differ from common methods in DM pre-processing. Rajaraman et al. ([Raja01], p. 104)

[20] Optical character recognition

and many other researchers arbitrarily define rules for feature selection according to this example: “All words appearing in less than 5% of the collection are removed, from each document; only the top n numbers of features based on TF-IDF ranking are picked”. A foundation for the chosen percentage as well as the value for “n” is not given. Therefore, the further results are not expected to be free of influence from this initial choice. Also, [Boll05] only generally describes this step with the words: “We then removed stop words (e.g., "the", "a", "I", "we", etc.)”, even though a common understanding of what a stop word is does not exist.

The following processing is proposed considering “the task of extracting trends or domain progress from textual data”. Task specific in this case is a time-based *horizontal* segmentation to bring the text collection in a timeline order. Here the decision of time granularity must be made. For the observation of knowledge domains (e.g., business informatics) a yearly segmentation appears to be useful. Otherwise seasonal components must be considered which would require spending effort on additional steps, but not necessarily raise the quality of the extracted results. Other domains may need different aggregation, e.g., for a company’s ad-hoc reporting a daily or weekly time granularity may be sufficient. The terminology “segments” for such time-ordered corpora is commonly used (see Rajaraman et al. [Raja01], p. 103).

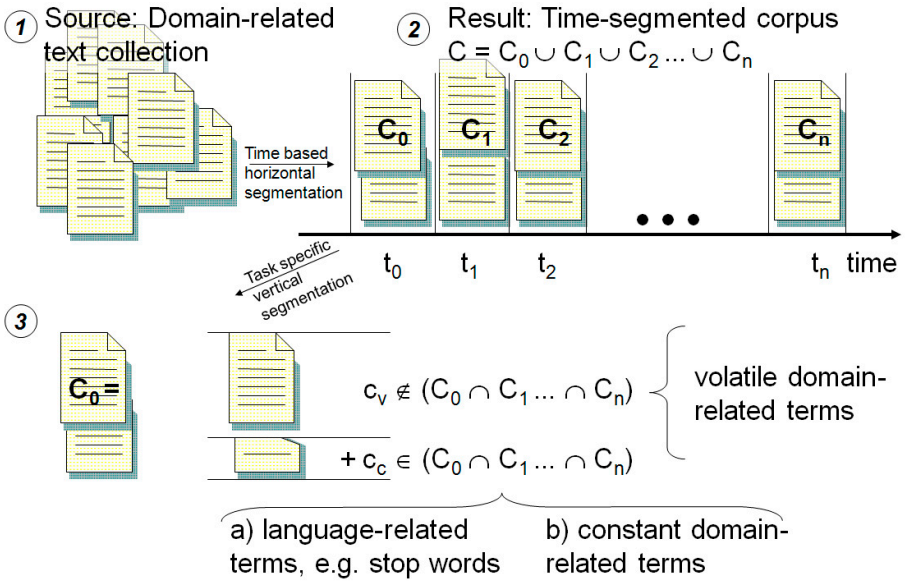


Fig. 24: Horizontal and vertical segmentation of a domain-related text collection

The corpus C consists of a number of time-sliced sub-corpora depending on the granularity of time (see No. 2 in Fig. 24). Horizontal segmentation here means a time-based horizontal segmentation. The decision of time granularity must be made. For the observation of knowledge domains (e.g., business informatics) a yearly segmentation appears to be useful. The segmentation granularity is domain dependent. Different domains may need different aggregation, e.g., for a company's ad-hoc reporting a daily or weekly time granularity may be sufficient. This horizontal segmentation results in time-sliced corpora (No. 2 in Fig. 24). A formal representation of this horizontal segmentation is given in 0:

$$[2] \quad \sum_{0..n} C_n$$

Now the *vertical* segmentation is applied to all segments and is exemplarily described. Each time segmented corpus part consists of terms that may also occur in other time segments ($c_c \in (C_0 \cap C_{1..} \cap C_n)$ and others that do not $c_v \notin (C_0 \cap C_{1..} \cap C_n)$. The whole corpus then is $C_{0...n} = c_v \notin (C_0 \cap C_{1..} \cap C_n) + c_c \in (C_0 \cap C_{1..} \cap C_n)$ (see No.3 in Fig. 24). The vertical segmentation represents 0:

$$[3] \quad C = C_V + C_C$$

C_C in this case represents language-related terms, the so-called "stop words" (functional words, verbs and nouns) but also constant domain-related terms, which belong to the basement of the certain domain. This can be the terms "calculate" or "algorithm" for business informatics or "transportation" and "sights" for the travel domain. C_V mainly contains the volatile domain-related terms that mark trends within the progress of a domain. C_c can therefore be

seen as an exactly matching task-specific generated stop-word list that is much less arbitrary than manual-produced stop-word lists.

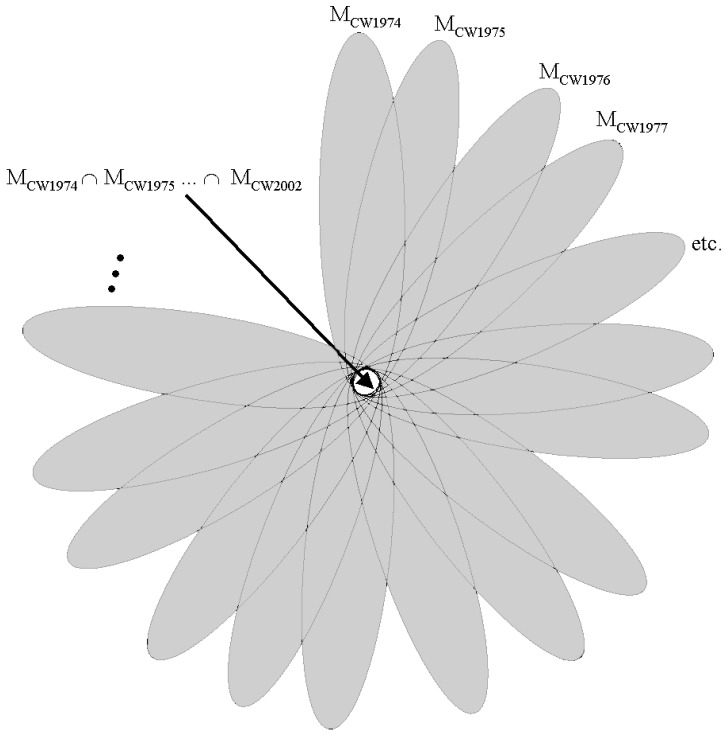


Fig. 25: C_V in a time-segmented corpus (marked grey)

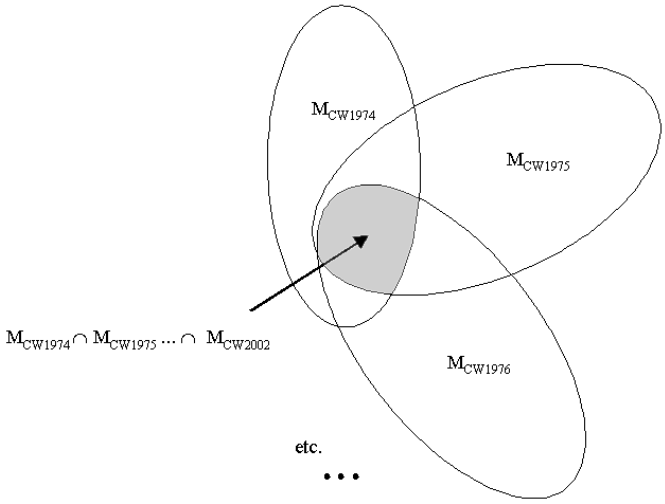


Fig. 26: Set of C_C in a time-segmented corpus (marked grey)

The different internal semantics of the corpus segments C_C and C_V must be considered within further knowledge-extraction processing.

3.5.2 Corpus measure based domain progress extraction paradigm

In this chapter the idea of using text collections for tracking domain knowledge, e.g., certain topics or terms and the dependencies between the pre-processing strategy and the extracted results are focused on. In focus is an end-to-end process that allows comparing knowledge which was extracted using different pre-processing strategies, applied to the source data from a knowledge worker perspective? The overall aim is to extract time-related knowledge from text collections and their aggregation to topics or major concepts. The process for that is oriented on the standard DM process consisting of the general steps of pre-processing, pre-filtering and corpus measure pattern recognition, data processing and data mining as well as domain knowledge interaction. This chapter will follow that main structure.

According to the research goal being analysed, if corpus measure-based processing can support TDM knowledge-extraction processes, then the general “corpus measure-based trend-extraction paradigm” is first introduced. This will be the basis for later evaluation scenarios. Every process step needs to be evaluated appropriately. All steps that have an input and a certain result, which may be determined by the process step, will have an individual evaluation discussion.

The TDM definition 0 contains the term “…previously unknown…” and in my opinion a very important precondition for the decision whether a procedure that is applied on textual data is information *retrieval* or information *generation*. From the opposite point of view, extracting only new things implies that *all previously known facts* must be considered *before* the generation of new information can begin. Stating this philosophic beginning in simpler terms, it means that all knowledge about the data must be used before the original data mining is applied. Projected on the DMP from [Fayy96] shown in Fig. 4, this must be done within the previous steps *Data Selection*, *Data Preprocessing* and *Transformation*. *Data Selection* will be discussed later. The focus here is on the two other steps. These steps are currently seen by most data-mining researchers only in a very data-quality-focused way where the task is simply to technically prepare the dataset for further steps. [Otte04], for

example, describes a wide range of possible conversions and transformations.

3.5.3 Corpus measure selection

A basic assumption is that the frequency of terms is positively correlated with their importance within a corpus (no articles et cetera, but nouns and names). Baayen et al. [Baay93] also assumes that the chances of once-used terms to be reused within texts are not independent to non-used terms. The randomness of the urn model and the randomness assumption is violated by the discourse structure of texts. To separate between terms that indicate a trend or hype and others that do not, requires that a threshold must be defined and made operable. The lack of classical tabular DM approaches is the bottom-up paradigm: After the data conversion focused pre-processing, every single data element (in the current objective: each term) is processed. From the bottom (the individual observation of each single feature value) these approaches generalize to form a rule that best fits the real world. The aim of this is to form a model of the real world by these rules found bottom-up. Benchmarked by the research objective of this dissertation – tracking domain progress and finding trends – the focus is on a highly aggregated perspective on the data. The measures used for tracking this progress on domain level must therefore be able to reflect this progress. Due to the fact that the data here is of textual source, measure candidates from textual processing research fields will be discussed later. In general, there are two classes of measures available, growth functions of corpus vocabulary, e.g., Orlov [Orlo83] developed a model of occurrence structures of vocabulary to learn more about typical term structures within corpora. Other research activities focus on corpus constants for inter-corpus comparison of corpora and corpus segments. Their possible task-related application in trend extraction is discussed in later chapters.

Statistical measures from corpus linguistics are potentially appropriate to measure qualities of corpora as a whole. Applications from corpus linguistics research include grammar and word use. Quantitative tasks are automatic

structure analysis and identification of contents (see [Lend98], p.14) that enable applications for semantic and style analysis. From CL the so-called measures of lexical richness and stylometrics are known[21] with focus on the language and word use. Simple measures from these fields will be introduced in this chapter which may be candidates for use as indicators of the importance of certain concepts or terms. In general within this research, appropriate measures must fit two conditions to support the measurement on corpus level:

a) They must be computable on corpus level (trivial)

b) There must be a transformation available to adopt results from corpus level to term level

The last condition is especially important to identify terms that lead to certain patterns on domain level. It must be possible to transform a certain measure value into a computable rule for a concrete decision, if a certain term is of high importance in reflection of semantics of a text or not. The selected terms constitute the set of the *potentially interesting* terms in the DMP that will be the basis for further analysis or interpretation. Based on the above-mentioned preconditions only vocabulary richness measures appear to be appropriate. In the following, a selection of known measures for vocabulary richness is presently introduced and an assessment regarding its suitability for measuring progress of domain is carried out. The statistical foundation of most measures is based on the urn model, which Yule first assumed in his research in word-distribution studies [Yule44]. The assumption is that words are used randomly and independently in texts. Transferred to the word-use case in texts, the use of a word can be modelled as the random selection of a marble from an urn. The urn typically contains a large number of marbles of various colours. Some colours appear on many marbles, others on just a few. In reality of course the probability of the use of a word in a text should rise if it is used for the first time in that text. This is to be expected because an author

[21] Research aim in this field is to learn more about the use and meaning of certain grammatical forms and word types. Researchers in this field are, for example, Michael Oakes or Douglas Biber. Please refer to their original research.

has a limited vocabulary and it is to be expected that the text is related to a certain topic. But this simplification is widely accepted and may work for most of the real-world text samples. Current research results draw this acceptance into doubt. E.g., Tweedie and Baayen added valuable research results to this field, e.g., with their 1998 issue “How Variable May A Constant Be? Measures of Lexical Richness in Perspective” [Baay98] to which this short introduction partly refers. Their aim was to analyse theoretically and empirically, whether a popular lexical richness measure is corpus-length-independent or not, motivated by the fact that many researchers create or use “constants” for the purpose of text characterization. They described a rich number of corpus measures for several purposes of which promising ones are selected for the current task. The subject of investigation here is not the quality of measure – being constant or not, being dependent or independent – but rather whether they help to measure vocabulary richness and possible existing preconditions. They must be taken into consideration in later steps of pre-processing. The most important implication from the work of Tweedie and Baayen in my research is: If simple quantitative measures are used then the precondition for comparable results is their application on corpora with equal corpus length.

Some popular stylistic measures that count word, sentences or text length are non-appropriate by definition, because they lack a direct linkage with certain terms. Other measures are indeed term related, but are not directly able to be aggregated at corpus level. Examples are:

- • Term Frequency Inverse Document Frequency
  which is defined as the number of times a term appears in a document multiplied by a monotone function of the inverse number of documents in which the term appears [Crou90].

  $$TF - IDF_T = \frac{F_T}{D_T} \quad [4]$$

  F_T is the total number of times a certain term occurs within the corpus. D_T is the number of documents the term occurs in. Different from some other sources here the former defined “term” was used instead of “word”.

Mertens used this approach intuitively while manually measuring the progress of the business informatics domain [Mert95]. This measure works on term level by definition and is therefore not a suitable measure for the meta level.

- Vocabulary size (V)
  The vocabulary size depends on the corpus length, N. As we read through a text, N increases from 1 to the total number of word tokens in the text. A word token is an instance of a particular word type. For instance, the preceding sentence contains two tokens of Type A. As the corpus length increases, the number of different word types encountered also increases, quickly at first then more slowly as additional text is read. The first panel of Figure 1 illustrates this functional dependence of the number of types on the number of tokens for Lewis Carroll's Alice's Adventures in Wonderland. The horizontal axis displays the corpus length in word tokens; the vertical axis shows the vocabulary size in word types. The second panel plots the growth rate of the vocabulary.

$$P(N) = \frac{V(1, N)}{N} \quad [5]$$

As a function of N (Good, 1953; Chitashvili and Baayen, 1993), where V (i,N) denotes the number of types occurring i times in the text at length N. The number of types occurring once, V (1;N) is generally referred to as the number of hapax-legomena. This plot highlights the diminishing rate at which the vocabulary increases through the text.

- Mean Word Frequency (MWF)

$$MWF(N) = \frac{N}{V(N)} \quad [6]$$

The MWF is dependent from the vocabulary and is the reciprocal of

- Type Token Ratio (TTR)

$$TTR(N) = \frac{V(N)}{N} \quad [7]$$

is defined as the ratio of different terms to their frequency of occurrence within a corpus [Lend98]. This measure then has a real corpus measure quality in relation to a certain domain. To measure progress at domain level based on the CL measure TTR a real corpus measure was defined, the

- Term-Repetition Quota (TRQ)

as a measure that is inherent to the term definition 0. The definition of a special measure was decided to make a distinction between the not completely identical context of both measures, TTR and TRQ.

[o] „The Term-Repetition Quota (TRQ) is defined as the ratio between all terms to all different terms occurring in a corpus.“

$$TRQ(N) = \frac{N}{V} \quad [8] \qquad [1; N]$$

In 0 is N the corpus length and V is the vocabulary (different terms) within the text. This definition is compatible with a Token per Type Ratio (TTR) on meta level. For later processing the projection of TRQ (meta level) on TTR (term level) will play an important role. TTR and the term frequency do have an equal meaning. It must to be taken under consideration that only a few (short) words or terms often occur in a corpus. More important words or terms are longer, but rare [Zipf49]. A usual method to consider this is to filter

very frequent, but not meaningful “stop words”[22] to focus the analysis on more important semantic terms.

The TRQ has a value of 1 if every term occurs once in the text or V is equal N. The opposite means that N is infinite and V not (is 1 at minimum). With TRQ there is a simple measure on meta level of the text available, which allows distinguishing between different texts. Due to TRQ’s close relationship with the TTR it is to be considered that TRQ is corpus-length dependent [Baay98]. Baayen found that the term distribution within texts is not even and therefore TRQ values can only be compared among each other, if they were derived from texts with an equal corpus length. The TRQ value is a mean average value of all TTR a time segmented corpus has.

The task-specific paradigm for mining trends from textual sources can be tailored as follows: If it is true that terms that are often repeated in a corpus, from which all stop words that have no meaning to the domain have been removed, then the TRQ has a median quality to a corpus and is an indicator of the importance of certain terms. In other words: The TRQ of a corpus divides this corpus into a set of terms that have a TTR that is below the TRQ value and another set that has a TTF that is higher than the TRQ value. Depending on the task a later processing can be done with this segmented corpus.

3.5.4 Discussion: Implications of TRQ value as threshold

In the meaning of Zipf’s “law” 0 that was criticized by several researchers after being published in 1949 [Zipf49], e.g., Herdan [Herd60], to be not a real law, offers a rough but empirically confirmed statement about the distribution of terms within texts.

[22] An approach of a task-specific definition for a stop-word list based on a semantic corpus segmentation is discussed later in this text.

$$[9]\quad f(w) = \frac{C}{r(w)^a}$$

This statement means that the often repeated terms do not have an important semantic for texts, because these terms are mostly articles, nouns and function words.

In 0, f (w) and r (w) are frequency and rank of word w and C are constants to be determined on the basis of the available data. Regarding the same, Zipf's Law is subject to several research discussions which not will be extended here. The lesson from this is only the fact that the distribution of terms within most real-world texts is approximated by formula 0. If this is applied to real texts, a similar graph of a hyperbolic distribution of terms like in Fig. 27 results if they are ranked according to the number of occurrences within the text. The corpus used in this approach is cleaned of such stop words, which results in the assumption that the concepts that are important for a certain domain are often repeated within domain-related corpora.

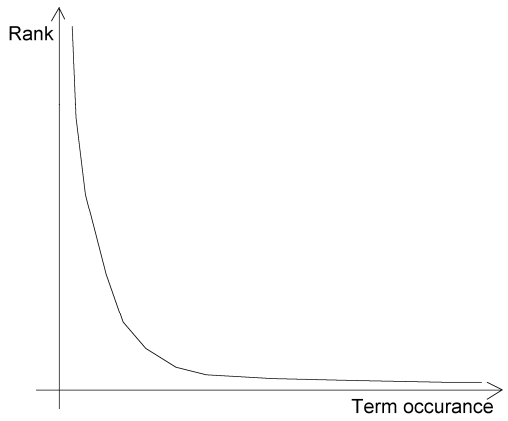


Fig. 27: Hyperbolic distribution of terms (schematic graph)

This assumption complies with other work, e.g., [Boll05] who combines this with the assumption that co-occurring terms are related. On this basis it is to be expected that the progress of a domain is reflected by the terms that are most frequently used. Topics and problems which involve a lot of human actors in discussions are made explicit in this way. The empirical finding of Zipf – that the more often repeated terms have a less semantic meaning – implies

that a filtering of such semantically irrelevant language belonging to terms should result in a corpus with a lower vocabulary growth function. Filtering of non-relevant terms lowers the vocabulary growth of distinct terms. New research actualised and extended Zipf's work, e.g., Baayen and Tweedie [Baay00] with their definition of mixture models for word-frequency distributions that covers distributions of terms coming from mixed sources. Here the assumption is made that the sources analysed are either homogeneous or mixed in a way that they can be treated as homogeneous.

TRQ_{AVG} is the mean value of TRQ for a certain corpus. The segmentation of the original source corpus C results in disjunctive sets of terms:

[10] $C = C_{TRQ}<TRQ_{AVG} \cup C_{TRQ}>TRQ_{AVG}$

Without any application of a DM method the corpus is now segmented task-specific and is ready for further steps. A detailed process description for tracking the domain knowledge over longer time periods follows in the next chapter. A detailed introduction of the measures used is given in Chapter 4.2.2.

3.6 Data processing and text (data) mining

Whereas this process step is the main focus of many TDM research projects, here only an exemplarily and simple statistical approach is used to evaluate the influence of pre-processing quality on extracted results in TDM.

As described in previous chapters the source data exists in yearly (horizontal) corpus segments (see 0). Each yearly segment consists of corpus-wide existing concepts and volatile concepts (see 0). The concepts were taken "as they appeared", but converted to a common ASCII format.

As an example of a simple DM method, statistical TRQ thresholds are applied on these prepared corpus segments that allow filtering significant, less frequently appearing terms than TRQ_{Mean}. The remaining terms are assigned

to taxonomies, which are built according to the procedure introduced in the next chapter.

3.6.1 Taxonomy construction

To represent semantics in a structured, accessible form is one of the major tasks in the approach introduced here. The terms that were clustered according to their persistence qualities and filtered by the help of statistical thresholds must be organized in a way that permits cognitive interaction. A logical structure is needed, therefore, that is capable of modelling simple semantic relationships as “is summarized by” to find more general concepts for grouping of terms.

For this analysis, taxonomy was manually built out of the corpus. The terms were assigned to dimensions, e.g., vendor, programming language and others. The built taxonomies are simple, directed graph representations of the semantic of the terms within the corpora. Of course within this step existing, externally defined and more complex ontologies could be integrated instead.

For both sources CW and AI1k different taxonomy creation processes were used in detail. For the taxonomy creation process expert domain knowledge was used, especially for the creation of domain-related concepts. Only terms according to definition 0 were considered.

CW

a) Order all terms according to their yearly frequency within the test set.
b) Delete all terms which appear fewer than 21 times in at least one yearly corpus segment.
c) Process next term.
d) If the term is domain specific and an appropriate dimension exists, assign this term to the dimension and proceed with c).
e) If the term is domain specific and no appropriate dimension exists, create an appropriate dimension and assign the term to this dimension, proceed with c).

f) Assign the term to the “ignore” dimension.

The restriction to consider only terms that appear at least 21 times in at least one yearly corpus segment may seem arbitrary. But compared to a yearly amount of around 50 issues of the “Computerwoche”, a term occurring fewer than 21 times cannot significantly dominate a yearly corpus segment.

Al1k

a) Process next term.

b) If the term is domain specific and an appropriate dimension exists, assign this term to the dimension and proceed with a).

c) If the term is domain specific and no appropriate dimension exists, create an appropriate dimension and assign the term to this dimension, proceed with a).

d) Assign the term to the “general” dimension.

Due to the fact that the Al1k corpus only contains 1,000 terms per yearly segment, it was not necessary to apply a selection of potentially significant terms like was necessary for the CW corpus.

3.6.2 Decomposition of constant domain-related and language-related terms

The decomposition of C_C into constant domain-related terms (C_{C_D}) and language-related terms (C_{C_L}) is a special challenge because prior knowledge was already applied to the source corpus. The further strategy certainly depends on the qualities of the elements of C_{C_D} and C_{C_L}.

The task here is to identify pure language-related terms to be able to separate them from the target data that is only domain related: $C_{C\_D} = C_C - C_{C\_L}$.

Because C_{C_L} contains terms that appear in all other time slices, these terms are not new to the language the corpus is written in. This simple observation may necessitate using a large dictionary of all grammatical forms to eliminate them from C_{C_L}. It is to be expected that the larger C is, the more C itself has

the quality of a large domain-related dictionary. A dictionary with abundant forms may be appropriate here.

If one filters additionally the persistent corpus elements C_C from a corpus, this has two implications:

1. The result set of terms is reduced by terms occurring in all periods (trivially). This means that terms, which describe constant circumstances in the business informatics, are no more available for further analysis.
2. If the TRQ time series is dependent substantially on terms from C_C, this is no longer recognizable.

These implications are logical and may seem obsolete to state here, but they are to be considered in evaluation process steps.

3.6.2.1 Discussion: Qualities of volatile domain-related terms

The larger C, the more C_C acts like a large dictionary and then $C_V = C - C_C$ results in a set of domain-only-related terms. But there might be bias due to the long timeframe under which the corpus was created. Usually 20 to 40 years have to be observed to see significant domain progress. Some domains, e.g., informatics, may need shorter observation periods; others may need longer, e.g., religion. But all corpora that are produced over longer time periods lack these points:

- Style inhomogeneity: Different authors tend to have their own writing style with different vocabulary and grammatical forms. But also a certain author may change his writing over time due to different corpus length [Zipf49], and learning effects [Grah00], [Hoov03].
- Progress in language: During a long observation period, new words and forms, as well as adoptions from other languages enhance the writing over time. All these new words do not belong to C_C per definition and therefore they add noise to C_V.

- Slow-changing semantics: In the 1980s, for example, the term "mailbox" was used as the term of an attainable dialling knot for the data interchange in telephone-based computer networks[23]. This term has been used since the mid-nineties for naming (e) mailboxes that store messages within the internet. From the beginning of the new century, as mobile communications technologies became ubiquitous, the term is used with the semantic meaning of message storage in telecommunications networks. Another example is the term “surfer”. This term arose in the seventies naturally in connection with the “water sports” domain. Since the mid-nineties a “surfer” predominantly is understood as a person who uses the WWW.

The non-occurrence of phrases is delimitation criterion as regards content or is a result of a change in context in the use of certain terms, which is recognizable only under inclusion of a domain expert with that. Remedy creates the inclusion of the use context, but automatic methods are rare.

3.6.3 TDM on segmented corpora based on TRQ threshold

For the filtering of uninteresting terms within the knowledge-extraction process a threshold based on TRQ measures for each yearly corpus segment is proposed here. For the calculation of lower and upper confidence intervals a time series of TRQ measure values is needed.

In general the following steps must be applied for the method introduced:

- Defining a window of *i* time periods for mean TRQ calculation
- Deciding whether a period is significantly different from previous periods; applying t-test on TRQ level
- Classification of dominant terms T: $F_T > TRQ_{Mean}$
- Semantic clustering of dominant terms by assignment to domain-specific taxonomy

[23] In Germany a technology called “BTX” was very popular for doing so.

The decision was taken to consider the last 10 values (i = 10) before the actual year. Width of window may be adjusted according to focus of analysis and corpus or domain specifics. For the first ten corpus segments constant values for the lower and the upper border were used, based on these ten input values.

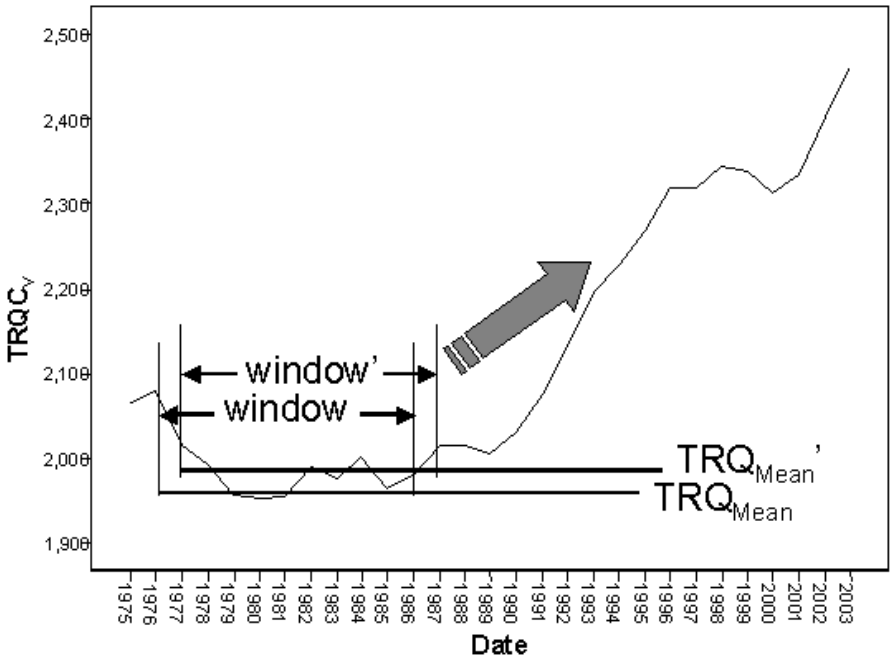


Fig. 28: Schematic graph: TRQ window for measure value tracking

Fig. 28 shows the main principle how the TRQ_{Mean} value is derived from the TRQ time series. With moving the window to window' … a time series for TRQ_{Mean} and TRQ_{Mean}' … is calculated from which it is possible to calculate the significant lower and upper borders. All terms that have a frequency out of this range are significant less[24] or higher[25] occurring within a corpus.

[24] below the lower border

[25] higher than the upper border

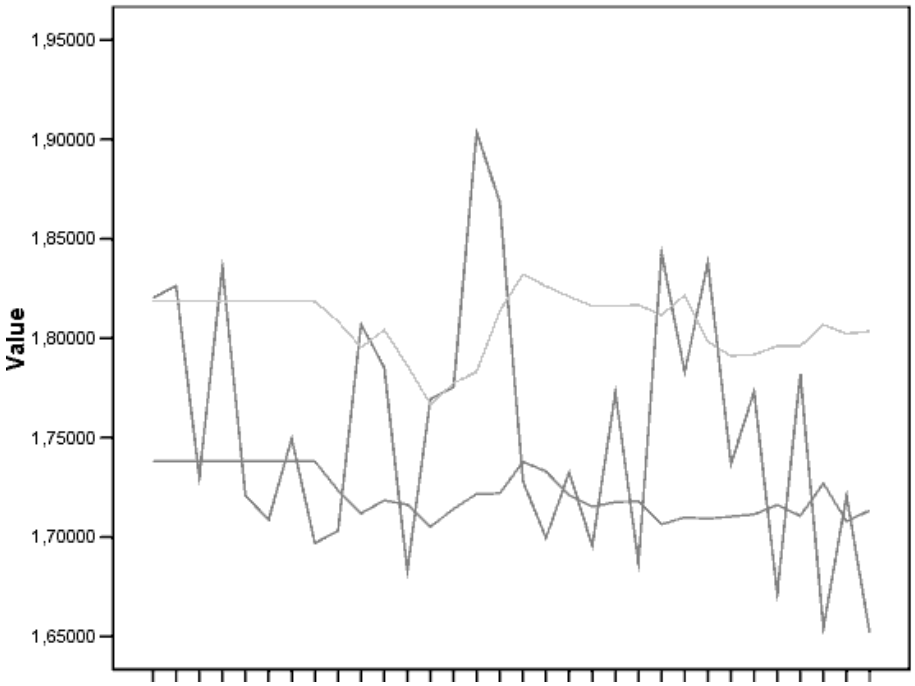


Fig. 29: TRQ plot with upper and lower confidence interval borders (example)

In the schematic plot in Fig. 29 shows an example of a frequency plot of a certain term that violates the lower and upper confidence interval borders for the TRQ_{Mean} value several times within the time series. This term is – in interpretation of the method here – dominant in some periods and not dominant in others.

3.7 Domain knowledge interaction

To be able to interact with it, knowledge must be represented explicitly. Theories on knowledge representation trace back to psychology and logic theory. Many theoretical models are available for description of concepts and their interaction[26]. Representation always depends on background knowledge and perspective of the individual person. In the field of computer sciences many research activities have been started in the last 15 to 20 years, which deal with knowledge representation and knowledge access, phenomenally expanded by the supplementation of the internet. First the focus was put on "webbing", distributed electronically sources to a large "world wide web"

[26] An overview is given by Gärdenfors, who introduces the theory of "Conceptual spaces" ([Gärd04], pp. 101).

source using hyper-linking techniques. As the first considerable markup language, the Standard Generalized Markup Language (SGML) arose, based on its predecessor, the Generalized Markup Language (GML). Charles F. Goldfarb mainly developed GML at IBM from 1969. The GML developers Goldfarb, Mosher and Lorie had been motivated by IBM to develop a description language for legal documents i.e. large textual documents. Two targets drove the development of GML: First, to be able to save, find, manage and publish documents automatically. Secondly, the communication ability of the text processing systems should be optimized under each other by the development of a structured generalized markup. With HTML (a standard, derived from SGML), a very popular standard was invented that allowed combining with hyper-linking to link such semi-structured documents and for representation of distributed knowledge. Today, SGML-based markup methods are available, e.g., defined in XML (see [Brad05]), for structuring text collections in databases in linguistics research. Two other research fields are present, which focus knowledge representation in a considerable level: Formal knowledge representation sciences and computational linguistics, which both developed special methods and tools.

Formal languages are used as logical representations of knowledge and for their physical processing and data exchange. Schelp [Sche99] proposes a conceptual modelling of multi-dimensional data structures that can be used as a basis for knowledge interaction. Based on such structures in databases various applications are possible to realize, e.g., applications for storing und querying historical texts (see Faulstich et al. [Faul05]). A popular concept for approaches that allow navigation within the logical structures modelled is Online Analytical Processing (OLAP). The data structure modelled within a database scheme allows for intuitively navigating along aggregation and disaggregation paths within so-called dimensions (the perspectives on data)[27]. In the approach introduced in this text a "Trend Landscape” metaphor[28] is proposed to be implemented as a technical solution. The idea is to model the structure of concepts according to their quantitative occurrence along a time-

[27] See Chamoni [Cham00], [Cham99a] and [Cham99b] for an introduction to main ideas and use cases of OLAP.

line with the ability to navigate along domain-specific aggregation and disaggregation paths. For this, a data model is defined that considers (simplified) the dimensions, term, date and time. The created data models are introduced in detail in Chapter 4.2.

3.7.1 Visualization approaches

Usually the visual sense is preferred as a basis for the transport of semantics from knowledge-extraction processes to humans. To lower the cognitive entrance barrier to complex patterns found by DM algorithms, various approaches within all sub-disciplines of DM were developed for establishing a visual representation of knowledge. General introductions are provided by Chen [Chen04] and Eppler [Eppl04]. Explorative data analysis is a subdiscipline in statistics and dates from the 1970s. Many of the methods used today for the visual exploration of multi-dimensional data have their roots in such statistical methods (see [Dege99c]). Many extensions of those methods have been made to adapt them to new tasks in actual scenarios of WWW and distributed computing with an enormous production of data.

Approaches which focus on the visualization of web usage, e.g., Behrendt [Bere03] with stratograms, build on the navigational behaviour of web surfers with consideration of the semantics behind their behaviour. Creators of ontologies support users to understand and use their ontologies within user tasks (see [Flui04]). Rauber et al. [Raub05] introduced an alternative approach based on self-organizing maps (SOM). This approach is iterative and allows evolutionary development maps with clusters of concepts. Without any dynamic in the presentation to the user it is not easy to visualize changes or trends over longer time periods. Two-dimensional concepts, e.g., "ThemeRiver" [Havr02] at which selected frequencies are represented in a kind of topic flow in the course of time, are conditionally limited concepts in their rep-

[28] see Fig. 2

resentation and interaction ability. Therefore, an OLAP-based approach with an additional time dimension is preferred.

A very comprehensive overview of historic and popular approaches of knowledge domain visualizing is provided by Börner et al. [Börn03]. The “User meta model”, their process flow for mapping knowledge domains (see Fig. 30), proposes a similar process to the TMF introduced in Fig. 10 in Chapter 3. The “User meta model” consists of the following steps: (1) data extraction, (2) definition of unit of analysis, (3) selection of measures, (4) calculation of a similarity between units, (5) ordination or the assignment of coordinates to each unit, and (6) use of the resulting visualization for analysis and interpretation.

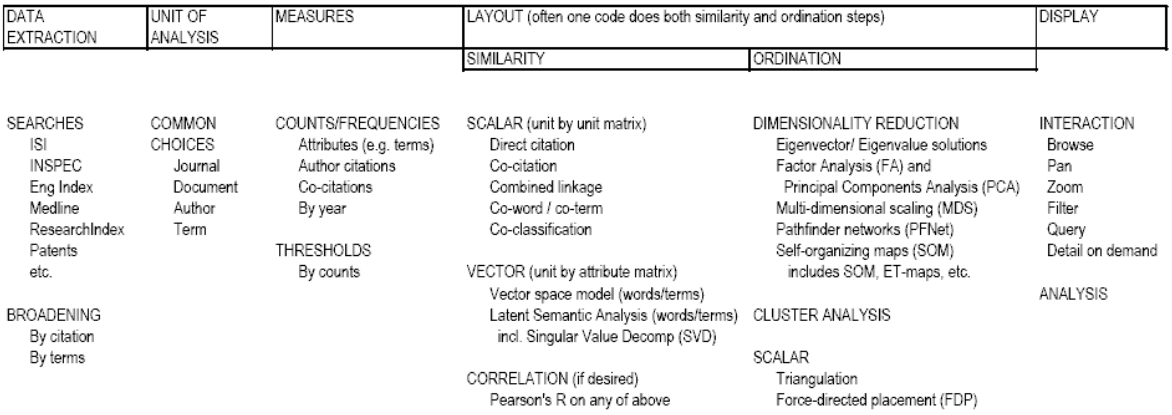


Fig. 30: Process flow for mapping knowledge domains (source [Börn03])

Within the “User meta model” visualization or display is the last step, the one that allows user-interaction. Within the approach introduced here the activities mentioned in Fig. 30 are extended with a capability to dynamically aggregate and disaggregate concepts through a dimension hierarchy. This permits a dynamical analysis according to the task the user is working on. An exemplary introduction of the possible interaction with the extracted knowledge is shown in Chapter 5.

4 Empirical results and evaluation

Evaluation is usually done by comparing precision 0 and recall 0 of different data-processing methods.

$$\text{Pr}\,ecision = \frac{true\_positives}{true\_positives + false\_positives} \quad [11]$$

$$\text{Re}\,call = \frac{true\_positives}{true\_positives + false\_negatives} \quad [12]$$

These measures can only be calculated if the number of *true positives* is known. The main challenge here is to define what “true positives” and “false positives” are. Another approach is to define a synthetic corpus that covers known topics. The benchmark here is to extract the topics included in the text collections as completely as possible. Fiscus et al. [Fisc02] describes the construction of appropriate TDT corpora for several generations of TDT tasks. Their corpora are based on manually annotated textual and acoustic corpora. In the opposite of such approaches, the corpora here will be non-synthetic, non-annotated empirically derived corpora. This implies that an absolute number of *true positives* will not be known. For this, evaluation measures derived in an optimal scenario will be declared as benchmark, to which all other scenarios will be compared. Further on the focus in this text is not to calculate the “real” true positives, but rather to compare measurable differences between texts that were differently pre-processed and the extracted knowledge based on these text samples. In ontology engineering equal data-quality aspects that lower the quality of extracted knowledge are to be considered like in general data processing (see Gómez-Pérez [Gome04a] for an overview)

In this current research project the analysis of the influence of different pre-processing strategies of source text collections on the extracted knowledge of domains is in focus. An adoption to other tasks can easily be made because neither the TMF components nor the TMF process is fixed.

The applied evaluation process will include the following tasks:

- Description of evaluation task (definition of qualitative criteria)
- Quantification of expected results by operable and measurable indicators
- Setup of an evaluation framework
- Data source selection and definition of appropriate test sets
- Extraction of quantitative measures from test sets
- Discussion of results

The next chapters follow this process chain. As the search for the “optimal” intensity of pre-processing is in focus within this research, appropriate test sets will be constructed for the following analyses.

4.1 Observed determining factors on knowledge extraction

For the execution of an analysis of unstructured text collections the transformation of not-quantitative data (different kinds of texts) into quantitative data will be necessary. This fundamental procedure for the quantitative text analysis among other things is described by Bailey (see [Bail78]) as the dominant aim. The evaluation is made using two general perspectives: The semantic and the quantitative perspective. The semantic is done as an expert evaluation of the extracted knowledge. The applied procedure remains constant regardless of the intensity of pre-processing the data.

For the empirical analysis, factors must be defined that will be adjusted in the analysis, and from that the determining power will be observed. With this varying of pre-processing based on the same source corpus, a measuring of robustness regarding the pre-processing is made possible. The following factors are adjusted between different sets of texts and the results of knowledge extraction then compared:

i. **Share of target data within analysed corpus (intensity of pre-processing)**

The intensity a corpus can be pre-processed is limited. Effort in this process step has decreasing returns. To have an idea of which effect different pre-processing strategies have is therefore economically interesting. In the following chapters an analysis of different intensities will be made, giving rough advice for an optimal pre-processing strategy at the end.

ii. **Language of origin**

When processing sources from different origins, the influence of different source languages on extracted knowledge will be in focus. The DM method used here, based on statistically significant TRQ thresholds, permits analysing the influence of single divergent TRQ values in certain time segments produced by different languages on extracted knowledge.

iii. **Corpus length**

Since Baayen and Tweedie showed the variability in so-called “corpus constants” [Baay98] a measurable influence of the corpus length on retrieved results in TRQ measures is to be expected as well as extracted knowledge. Corpus length is observed with focus on recognizable influence on extracted knowledge. Defining a “low border” of affordable corpus length with the use of the method is also carried out here.

iv. **Knowledge domain**

Methods in DM tend to “overfitting”: After a number of iterations of method adjusting according to a defined test set, the method extracts exactly the knowledge patterns that are within this certain test set. The application of two completely different test sets from different knowledge domains will give the chance for testing, whether the proposed method is capable of being successfully applied on different kinds of domains or not.

v. **Kind of document source (WWW vs. scanned texts)**

DM methods can focus certain kinds of data, e.g., click stream analysis mostly relys on server log files. The method used here is applied to data that was originally downloaded from the WWW (the CW corpus) and from scanned texts (Allianz management reports).

For the evaluation process the TMF (see Chapter 0) with its components is used. The canon of processing that is applied to each textual data source consists of:

a) Term-wise aggregation of weekly issues (Fig. 24, No.1) to yearly corpus segments (Fig. 24, No.2: horizontal segmentation)
b) Vertical segmentation of terms (Fig. 24, No.3) according to their occurrence across the yearly segments
c) Extraction of corpus metadata (number of types and tokens, calculation of simple TRQ and significant threshold values) for each segment, calculating statistics for corpus comparison
d) Multi-dimensional structuring of texts by applying different types of domain-related taxonomy to the test sets
e) Using TRQ as threshold for concept extraction
f) Benchmarking of each different pre-processed corpus on the basis of expected and extracted concepts (expert evaluation)

The steps mentioned above are applied to the text corpora. For a standardized process, independent from source data qualities and origin, a common multi-dimensional modelling procedure was applied. This will be introduced in the next chapter.

4.2 Data models

The corpora were converted into multi-dimensional views based on frequency lists of terms. Dimensions were defined as alternative directed graphs from certain taxonomy that allow desegregations down to term level.

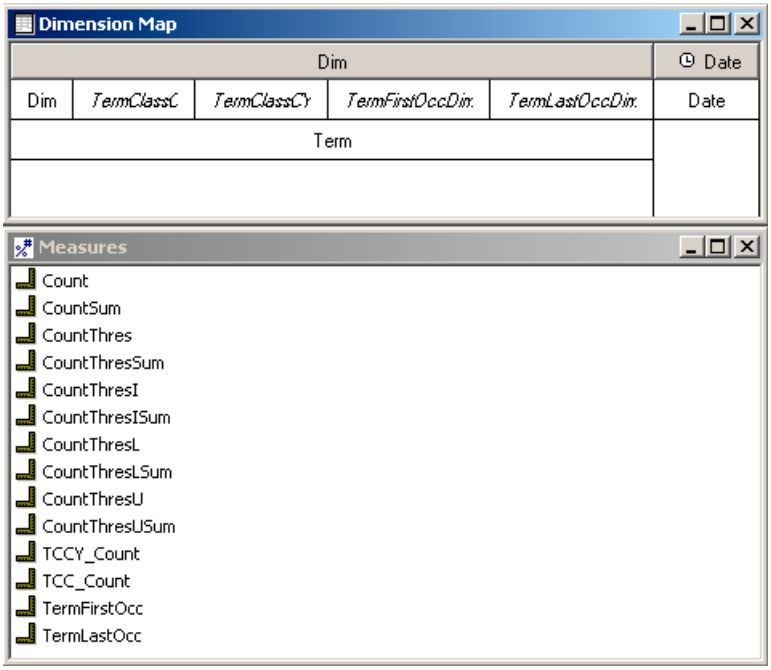


Fig. 31: Multi-dimensional data model with alternative desegregation paths (example)

Every time that more than one taxonomy was applied, the taxonomies are combined in an intersection mode so that the less-covering taxonomy limits the terms that appear within the whole multi-dimensional model. This multi-dimensional modelling allows the application of different perspectives on the same data. On the one hand, quantitative analysis is conducted using several statistical procedures while, on the other hand, threshold functions on the basis of TRQ measures support the extraction of terms and aggregated topics. The focus of the following chapters will be an introduction of the used dimensions and measures.

4.2.1 Dimensions

In

Table 4 all dimensions are introduced that were used in the different data models.

Table 4 Dimensions / taxonomies used in the models

Dimension/ taxonomy	Appears in models as	Description	Types
Dim (CW data models)	Dim	Taxonomy with all terms assigned to, which appear at least 21 times in one yearly corpus segment (for details refer to Appendix)	3,581 domain specific; 6,449 stop words (filtered)
Dim (AI1k data models)	Dim	Taxonomy with all terms assigned to, which appear at least 21 times in one yearly corpus segment (for details refer to Appendix).	1,241 domain specific; 233 stop words (filtered)
Dim_Mertens (CW data models)	Dim_Mertens	Taxonomy, constructed based on selected topics that were tracked in [Mert95]	10
TermClassC	TCC	21 categories of terms that appear 1 to 21 times within the whole corpus (Categories 1-20 *may* be empty, if one of the "Dim" dimensions is used parallel in a model, acts as QA dimension then)	Dependent on source corpus
TermClassCY	TCCY	21 categories of terms that appear 1 to 21 times within a certain yearly corpus segment (categories 1-20 *must* be empty, if one of the "Dim" dimensions is used parallel in a model, acts as QA dimension then)	Dependent on source corpus
TermFirstOccDim	TermFirstOccDim	Year of initial appearance of a certain term	Dependent on source corpus
TermLastOccDim	TermLastOccDim	Year of last appearance of a certain term	Dependent on source corpus
CC_Dim	CC_Dim	Constant terms of AI1k corpus, assigned to linguistic classes	46
Date	Date	Time dimension, contains year of corpus segment	Dependent on source corpus

The Dim_Mertens taxonomy (see Table 5), constructed on the basis of selected topics that were tracked in [Mert95], is also an example of how a hierarchical structure can be used for summarizing different linguistic terms into one common category. This is done here with the concept "BTX" aggregating the different written terms "BTX" and "Btx".

Table 5: Dim_Mertens taxonomy

Dim	Term
BTX	BTX
BTX	Btx
CIM	CIM
Datenschutz	Datenschutz
Dezentralisierung	Client-Server
Dezentralisierung	Dezentralisierung
Dezentralisierung	Downsizing
EDI	EDI
KI	KI
Outsourcing	Outsourcing

The dimensions were arranged in different combinations. Every dimension may have its own stop-word list, realized with marking this certain category within a dimension with a red cross (see Fig. 32). If a certain term belongs to such a deleted category within a data model, this will lead to the absence of this term in the resulting data set, even if this term belongs to one of the other dimensions as a usual category.

Both “Dim” dimensions (CW and Al1k) were built under consideration of all terms that occur at least 21 times in a certain year. This threshold was chosen due to the fact that, especially if looking at the weekly publication “Computerwoche”, a concept or term could not dominate a yearly time period if it only appears much less than every second issue of that publication.

The Allianz and the CW “Dim” dimensions were built on the basis of the largest available domain corpus (Al1k and CW_{5k}) with manual assignment of every term to a taxonomy dimension. The creation of a new domain was triggered by the simple rule that a new dimension was created if the existing did not fit. A certain term was assigned to only one dimension structure level to avoid double counts.

The dimension structures are explained in Table 6 and Table 7.

Table 6: Description of dimension structure of CW_{5k} corpus

Dimension structure	Description
Business	General business-related terms, e.g., “revenue”
Currency	Abbreviations and long descriptions of currencies
Customer	Companies and organizations that predominately consume IT products and services
Economy	Economical terms above company level
Event	E.g., fairs
Geography	Countries, towns, landscapes
Ignore	Stop words
Institute	E.g., research organizations
IT	General IT domain-related terms
ITProduct	IT domain-related products with a certain trade name
Name	Person names
Norm	Technical norms, also abbreviated
OS	Operating system names
Performance	IT domain-related terms that describe performance quantitative or qualitative
Profession	Professions and roles within organizations and companies
ProgLanguage	Programming language

Dimension structure	Description
Science	Research-related terms
SocialFramework	Terms that are related to the social background, e.g., “Politik”
Vendor	Companies and organizations that predominately offer IT products and services

Table 7: Description of dimension structure of AI1k corpus

Dimension structure	Description
BusinessTerm	General business-related terms, e.g., “revenue”
Company	Companies and organization names
Currency	Abbreviations and long descriptions of currencies
General	Stop words
Geography	Countries, towns
InsuranceTerm	Insurance domain-related terms
Name	Person names

Based on the AI1k corpus a linguistic taxonomy was built[29] that allows simple linguistic analysis. For the assigning of terms to the dimension structure levels more than one option was given in some cases. Here, one of these options was chosen without intensive analysis. This was done in this work due to the focus on DM and not in CL.

Table 8: Description of dimension structure of CC_Dim

Dim
Article
Conjunction
Particle
Preposition
Pronom
Verb

If combined the dimensions introduced earlier constitute a multi-dimensional data model. Beginning at the start node the dimensions support alternative disaggregation paths through the dimension structure down to term level.

[29] For a complete overview of assigned terms refer to Appendix.

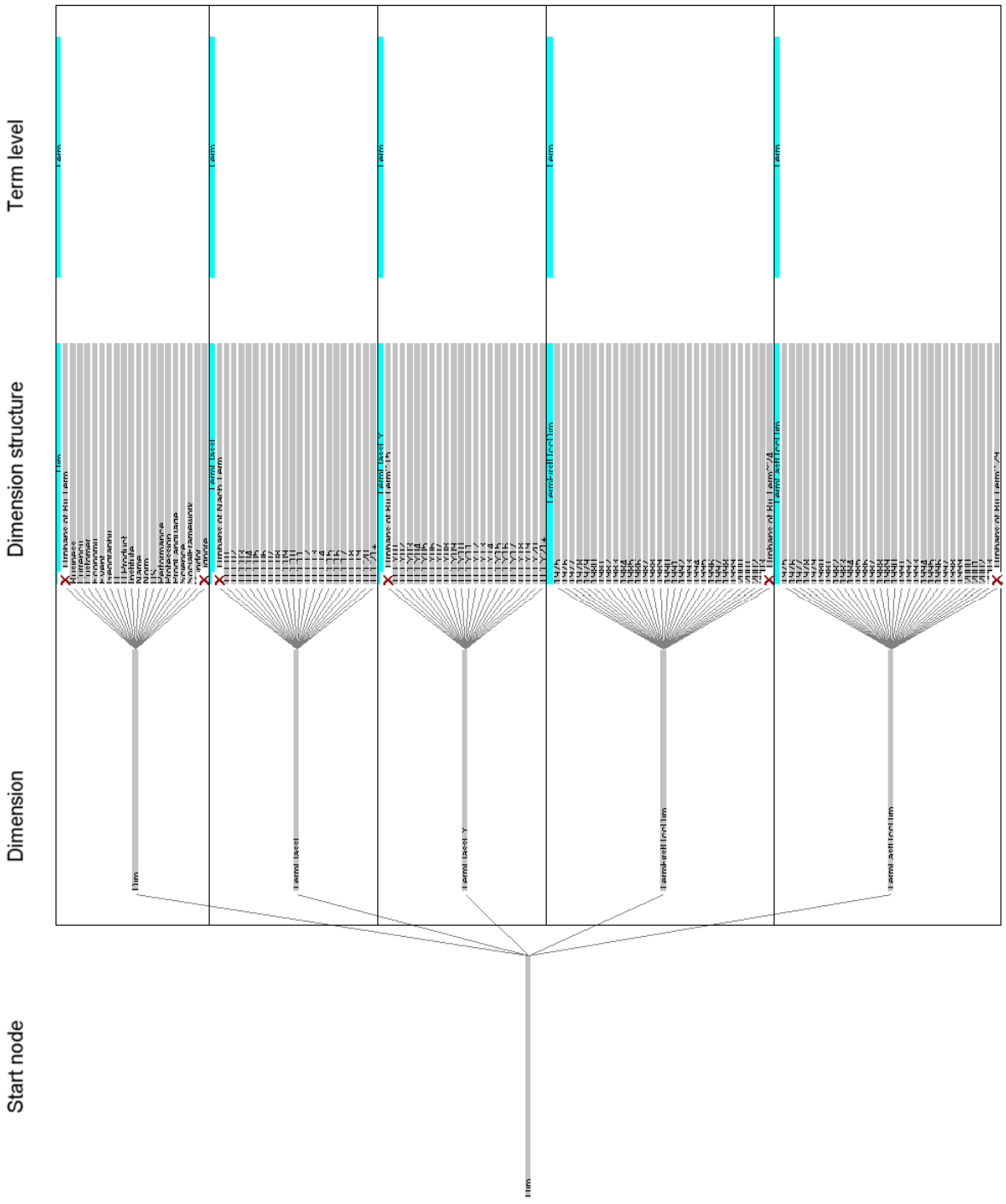


Fig. 32: Taxonomies as directed graphs

The terms assigned to the directed graph structure (see Fig. 32) allow an analysis from various perspectives, linguistic and domain specific. In the following analysis, these combinations of taxonomies were used:

Table 9: Dimension / taxonomy combinations used in the models

Suffix in model names	Dimension / taxonomy combination						Used with corpus segments
No suffix	Dim	TermClassC	TermClassCY	TermFirstOc-cDim	TermLastOc-cDim	Date	C, C_C, C_V
CC_Dim_alone	CC_Dim	TermClassC	TermClassCY	TermFirstOc-cDim	TermLastOc-cDim	Date	C, C_C[30]
Dim_Mertens	Dim_Mertens	TermClassC	TermClassCY	TermFirstOc-cDim	TermLastOc-cDim	Date	C, C_C, C_V
Dim_Mertens_alone	Dim_Mertens_alone	TermClassC	TermClassCY	TermFirstOc-cDim	TermLastOc-cDim	Date	C, C_C, C_V

4.2.2 Measures

The measures that were defined have quantitative (e.g., “Count” and “Count-Sum” in Table 10) or qualitative character (e.g., “Thres” as threshold for term filtering). With defining a threshold the input data sets were divided into a (not predominant) set of terms that appeared less than the average value of TRQ and another set that contains the other terms.

Table 10: Measures used in the models

Measure	Description	Range	Aggregation
Count	Number of token	[0;∞]	Average
CountSum	Number of token	[0;∞]	Sum
Thres	Number of token with TRQ>TRQ_{AVG}	[0;∞]	Average
ThresSum	Number of token with TRQ>TRQ_{AVG}	[0;∞]	Sum
ThresI	Indicator for significant deviations from 95% confidence interval of TRQ, based on a *t*-test of a ten-year TRQ window	[-1;1]	Average
ThresISum	Indicator for significant deviations from 95% confidence interval of TRQ, based on a *t*-test of a ten-year TRQ window	[-∞;∞]	Sum
ThresL	Number of token with TRQ>TRQ_{AVG} on the basis of the lower confidence interval border of a *t*-test of a ten-year TRQ window	[0;∞]	Average
ThresLSum	Number of token with TRQ>TRQ_{AVG} on the basis of the lower confidence interval border of a *t*-test of a ten-year TRQ window	[0;∞]	Sum
ThresU	Number of token with TRQ>TRQ_{AVG} on the basis of the upper confidence interval border of a *t*-test of a ten-year	[0;∞]	Average

[30] To combine constant term taxonomy with volatile data from the same basis corpus is not useful by definition and results in no matching assignment.

Measure	Description	Range	Aggregation
	TRQ window		
ThresUSum	Number of token with TRQ>TRQ_{AVG} on the basis of the upper confidence interval border of a *t*-test of a ten-year TRQ window	[0;∞]	Sum
TCCY_Count	Number of tokens of a certain concept or term within a yearly corpus segment	[0;∞]	Minimum
TCC_Count	Number of tokens of a certain concept or term within the whole corpus	[0;∞]	Minimum
TermFirstOcc	Year of initial appearance of a certain concept or term	dependent on source corpus	Minimum
TermLastOcc	Year of last appearance of a certain concept or term	dependent on source corpus	Maximum

Switching between the different measures automatically adjusts the number of terms within the result set and updates the presented view on the data sets. This filtering will be used for tracking terms from different perspectives and with different purposes of analysis in Chapter 5ff.

4.3 Evaluating the impact of intensity of pre-processing

If a knowledge worker is starting to analyse a large text collection, with the focus on a certain knowledge domain, it is not to be expected that he has knowledge of the share of data that is knowledge-domain related. Even if he had an idea of this share (e.g., by knowledge of general structure or formatting), the second problem is to identify the related text passages. In a real-world scenario it is seldom to be expected that the optimal situation (the whole data set) is domain-related.

Based on this assumption different data sets were prepared, which represent the data basis of the method evaluation regarding various semantic perspectives.

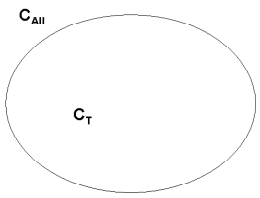
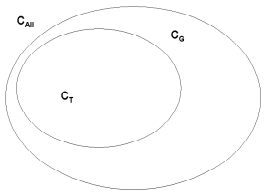
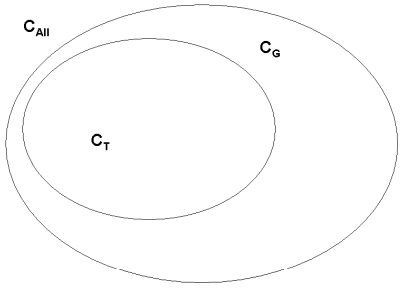
Let the whole corpus be:

$$\textit{[a]}\ C_{All} = \sum_{1...n} C_T \cup C_G\ ,$$

where C_T contains the target corpus data (consisting of task-specific semantic-relevant terms). Other terms are not relevant and constitute the non-target "garbage" set of non-relevant terms C_G. To have ex-ante knowledge regard-

ing the quantitative relation between C_T and C_{All} is seldom the case. A more realistic scenario is that the researcher considers the corpus length dependency of quantitative measures and tries to avoid bias by the selection of even segments of the corpus regardless of the share of target data. The following general, theoretical corpus types can be defined as follows:

Table 3: Theoretical corpus types

Corpus type		Venn diagram of yearly corpus segments	Description	Assumptions/ Remarks
n	net		Pure target data quantitative normalized	All non-domain-related terms were eliminated in pre-processing A human-readable pure ASCII text collection
bn	gross, normed		Corpus type b, but quantitative normalized token per yearly segment)	Due to a lack of identification of target data, pre-processed corpus also contains terms that belong to C_G C remains constant, C_G and C_T are volatile pair wise and between yearly segments
b	gross		Corpus type n + all non-target data formatting etc.[31]	Previously eliminated formatting etc. is added again C_T remains constant, C_G and as result, also C is not constant over the yearly segments

To find corpora mostly of type “bn” or “b” is to be expected in real case scenarios (see Fig. 33), which inherent a volatile share of target data combined with non-relevant content. In general the aim of pre-processing in data mining

[31] The original downloaded HTML files that contain the target content in a large surrounding of tags, advertisements and external links.

should be to identify the type of corpus and to apply appropriate pre-processing steps to transform such corpora into pre-processed corpora of type “n”.

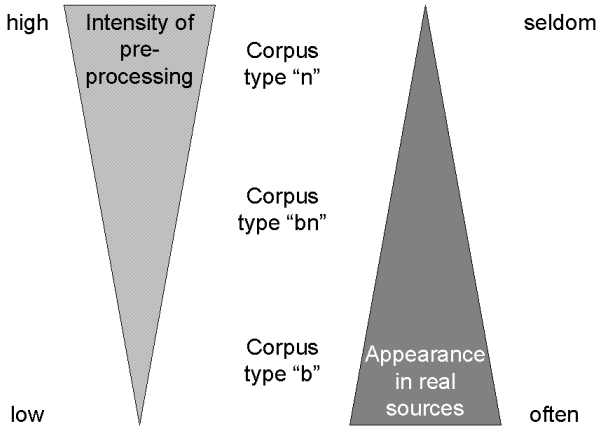


Fig. 33: Corpus types and their appearance in real sources

The data sets introduced in the following chapter were defined according to the idea of real-life scenarios based on different strategies of data preparation.

The following test sets were used:

Table 11: Overview on prepared test sets

Test set	Technical name (Prefix)	Corpus Type	Knowledge Domain	Source	Language of origin	Description (yearly segments)
CW_{5k}	CW5K	n	Business Informatics/ Information Technology	WWW	German	500,000 tokens overall per year, no C_G
CW_{5kb}	CW5KB	b	Business Informatics/ Information Technology	WWW	German	500,000 tokens C_T, surrounded by C_G, no “German umlaut” conversion into ASCII
CW_{5kbu}	CW5KBU	b	Business Informatics/ Information Technology	WWW	German	500,000 tokens C_T, surrounded by C_G, with “German umlaut” conversion into ASCII

Test set	Technical name (Prefix)	Corpus Type	Knowledge Domain	Source	Language of origin	Description (yearly segments)
CW_{5kbun}	CW5KBUN	bn	Business Informatics/ Information Technology	WWW	German	500,000 tokens C_T, surrounded by C_G, with "German umlaut" conversion into ASCII
CW_{5kbun2}	CW5KBUN2	bn	Business Informatics/ Information Technology	WWW	German	500,000 tokens C_T, surrounded by C_G, with "German umlaut" conversion into ASCII
CW_{1k}	CW1K	n	Business Informatics/ Information Technology	WWW	German	Subset CW_{5k}: 100,000 tokens overall per year, no C_G
CW_{1kb}	CW1KB	b	Business Informatics/ Information Technology	WWW	German	100,000 tokens C_T, surrounded by C_G, no "German umlaut" conversion into ASCII
CW_{1kbu}	CW1KBU	b	Business Informatics/ Information Technology	WWW	German	100,000 tokens C_T, surrounded by C_G, with "German umlaut" conversion into ASCII
$AI1k_{S1}$	AI1kS1	n	Insurance	OCR	German, English	1,000 tokens overall per year, no C_G
$AI1k_{S2}$	AI1kS2	n	Insurance	OCR	German, English	1,000 tokens overall per year, no C_G

Linguistic metadata of the corpora are documented in Table 12 (derived from the corpora stored within database tables):

Table 12: Corpus test sets metadata

Data Set	C (Token)	C (Types)	C_T (Token)	C_T (Types)	C_C (Token)	C_C (Types)	C_V (Token)	C_V (Types)
CW_{5k}	14,502,249	623,158	14,502,249	623,158	11,246,759	6,196	3,255,490	616,962
CW_{5kb}	432,966,111	775,309	14,502,249	623,158	390,393,405	5,800	42,572,706	769,509
CW_{5kbu}	430,697,494	781,258	14,502,249	623,158	389,923,506	6,765	40,773,988	774,493
CW_{5kbun}	14,499,822	86,530	<100%	<100%	12,783,666	817	1,716,156	85,713
CW_{5kbun2}	14,499,871	84,929	<100%	<100%	12,806,754	831	1,693,117	84,098
CW_{1k}	2,901,069	213,001	2,901,069	213,001	1,944,861	1,838	956,208	211,163
CW_{1kb}	87,749,805	259,792	2,901,069	213,001	78,171,046	2,102	9,578,759	257,690
CW_{1kbu}	87,297,586	261749	2,901,069	213,001	78,173,390	2,346	9,124,196	259,403

Data Set	C (Token)	C (Types)	C_T (Token)	C_T (Types)	C_C (Token)	C_C (Types)	C_V (Token)	C_V (Types)
$Al1k_{S1}$	31,967	7,609	31,967	7,609	8,328	22	23,639	7,587
$Al1k_{S2}$	31,948	7,642	31,948	7,642	8,100	21	23,848	7,621

The third and fourth columns “C (Types)” and “C_T (Token)” documenting the extreme dependency of the share of target data contained within the selected text collections on the pre-processing strategy. It is important to remember that the three different corpora (first three rows) contain exactly the same number of target data (see Table 12). Without any application of quantitative methods it can be seen that the non-content data dominates the data sets CW_{5kb} and CW_{5kbu}. Even after filtering the terms belonging to C_C in Data Set CW_{5kb} and CW_{5kbu} the target data C_V with a total of 3.2 million token in Data Set A is mixed with approx. 40 million non-target data in both cases. A remarkable result is that a very rough data preparation (German umlaut conversion) applied to Data Set CW_{5kbu} led to nearly the same target data amount as achieved in data set CW_{5kb}. The semantic differences will be analysed in the next chapter. The analysis is done both on term level and topic level based on a simple term dimension of assigning taxonomy for topic aggregation. Simple statistics for each specific corpus is shown in Table 45 (see Appendix).

The “optimal” CW_{5k} corpus, with an even yearly amount of about 500,000 tokens, without any non-domain-related terms, has relatively a low range and standard deviation and a moderate negative skewness and kurtosis.

The moderate negative skewness and kurtosis can also be found in CW_{1k} data set, but not in $Al1k_{S1}$ and $Al1k_{S2}$ test sets (see Table 47 in Appendix) with their limited number of terms.

The value in Column “Sum” in Table 45 to Table 47 (see Appendix) is equal to the value of column “C” in Table 12. Minimal differences may occur due to data conversion between the database and statistics program. In the following chapters the different corpus data sets introduced in Table 12 are analysed in detail regarding their statistical corpus measure qualities.

Results from corpus statistics summary
• *A low pre-processing intensity of corpora lowers the share of target data.* • *Filtering of the intersection data set of all time segments leads to a higher share of target data, but lower and more volatile distributed compared with optimal pre-processed data sets.* • *Even a more sophisticated processing (e.g., German umlaut conversion into an ASCII representation) applied to low-intensity pre-processed data sets does not raise the absolute number of target data contained in the test sets.*

4.3.1 Corpus type n (high pre-processing intensity)

Best results in TDM are to be expected when using source data, where $C_G \to 0$ (the optimal case). Within the following chapters corpora are analysed, with $C_G = 0$.

4.3.1.1 Statistical analysis of type n corpora

CW_{5k} is expected to be the best basis of corpus segments introduced in Table 11 due to the fact that it is not mixed with garbage data and is an even collection of pure domain-related terms. Therefore, CW_{5k} is used as a benchmark for all later analysis of prepared corpora from CW.

4.3.1.1.1 Descriptive statistics of CW corpus test set CW_{5k}

An overview of statistical qualities of the corpus C, the corpus segments C_C and C_V based on TRQ measures is documented in Table 48 (see Appendix).

The absolute value of TRQ is high within C_C and low within C_V. This is also the case for mean and standard deviation. The skewness is positive for C and C_V, but negative for C_C. The value of kurtosis is only positive for C, which

allows concluding that the distribution of values varies between the different vertical corpus segments.

Table 13: Correlations between corpus segments based on TRQ measure

Correlations

		Cw5kCRep Quo	Cw5kCcRep Quo	Cw5kCvRep Quo
Cw5kCRepQuo	Pearson Correlation	1	-,623**	,881**
	Sig. (2-tailed)		,000	,000
	Sum of Squares and Cross-products	2,299	-4,218	1,157
	Covariance	,082	-,151	,041
	N	29	29	29
Cw5kCcRepQuo	Pearson Correlation	-,623**	1	-,918**
	Sig. (2-tailed)	,000		,000
	Sum of Squares and Cross-products	-4,218	19,934	-3,551
	Covariance	-,151	,712	-,127
	N	29	29	29
Cw5kCvRepQuo	Pearson Correlation	,881**	-,918**	1
	Sig. (2-tailed)	,000	,000	
	Sum of Squares and Cross-products	1,157	-3,551	,750
	Covariance	,041	-,127	,027
	N	29	29	29

**. Correlation is significant at the 0.01 level (2-tailed).

In Table 13 a significant positive correlation can be found (based on TRQ measure) between C and C_V and significant negative correlations between C and C_C as well as between both corpus segments C_V and C_C. The segmentation of C into a constant and volatile segment led to statistically significant different corpus segments (99% confidence interval level).

4.3.1.1.2 Descriptive statistics of CW corpus test set CW_{1k}

An overview of statistical qualities of the corpus C, the corpus segments C_C and C_V based on TRQ measures is documented in Table 49 (see Appendix). The absolute value of TRQ is again high within C_C and low within C_V. This is also the case for mean and standard deviation.

The skewness is positive for C and C_V, but negative for C_C. The values of kurtosis are all negative; the values of skewness vary between the segments C, C_C and C_V.

Table 14: Correlations between corpus segments based on TRQ measure

Correlations

		Cw1kCRepQuo	Cw1kCcRepQuo	Cw1kCvRepQuo
Cw1kCRepQuo	Pearson Correlation	1	,033	,722**
	Sig. (2-tailed)		,865	,000
	Sum of Squares and Cross-products	,352	,043	,137
	Covariance	,013	,002	,005
	N	29	29	29
Cw1kCcRepQuo	Pearson Correlation	,033	1	-,665**
	Sig. (2-tailed)	,865		,000
	Sum of Squares and Cross-products	,043	4,707	-,460
	Covariance	,002	,168	-,016
	N	29	29	29
Cw1kCvRepQuo	Pearson Correlation	,722**	-,665**	1
	Sig. (2-tailed)	,000	,000	
	Sum of Squares and Cross-products	,137	-,460	,102
	Covariance	,005	-,016	,004
	N	29	29	29

**. Correlation is significant at the 0.01 level (2-tailed).

In Table 14 a significant positive correlation can again be found (based on TRQ measure) between C and C_V but no correlation between C and C_C. The significant negative correlation between corpus segments C_V and C_C is present as in CW_{5k}, but with a lower absolute value. The segmentation of C into a constant and a volatile segment led to statistically significant different corpus segments (99% confidence interval level).

4.3.1.1.3 Descriptive statistics of Allianz corpus test sets $Al1k_{S1}$ and $Al1k_{S2}$

The results of the analysis of test sets $Al1k_{S1}$ and $Al1k_{S2}$ are presented together in this chapter. An overview on statistical qualities of the corpus C, the corpus segments C_C and C_V based on TRQ measures is documented in Ta-

ble 50 and Table 51 (see Appendix). The absolute value of TRQ is high within C_C and low within C_V. This is also the case for mean and standard deviation. Skewness and kurtosis do not show comparable values in both test sets.

Table 15: Correlations between corpus segments based on TRQ measure

Correlations

		AI100S1RepQuo	AI100S1CcRepQuo	AI100S1CvRepQuo
AI100S1RepQuo	Pearson Correlation	1	,523**	,812**
	Sig. (2-tailed)		,001	,000
	Sum of Squares and Cross-products	,145	,868	,081
	Covariance	,004	,023	,002
	N	39	39	39
AI100S1CcRepQuo	Pearson Correlation	,523**	1	-,070
	Sig. (2-tailed)	,001		,672
	Sum of Squares and Cross-products	,868	19,074	-,080
	Covariance	,023	,502	-,002
	N	39	39	39
AI100S1CvRepQuo	Pearson Correlation	,812**	-,070	1
	Sig. (2-tailed)	,000	,672	
	Sum of Squares and Cross-products	,081	-,080	,069
	Covariance	,002	-,002	,002
	N	39	39	39

**. Correlation is significant at the 0.01 level (2-tailed).

Table 16: Correlations between corpus segments based on TRQ measure

Correlations

		AI100S2CRepQuo	AI100S2CcRepQuo	AI100S2CvRepQuo
AI100S2CRepQuo	Pearson Correlation	1	,462**	,770**
	Sig. (2-tailed)		,003	,000
	Sum of Squares and Cross-products	,146	,890	,079
	Covariance	,004	,023	,002
	N	39	39	39
AI100S2CcRepQuo	Pearson Correlation	,462**	1	-,207
	Sig. (2-tailed)	,003		,205
	Sum of Squares and Cross-products	,890	25,442	-,283
	Covariance	,023	,670	-,007
	N	39	39	39
AI100S2CvRepQuo	Pearson Correlation	,770**	-,207	1
	Sig. (2-tailed)	,000	,205	
	Sum of Squares and Cross-products	,079	-,283	,073
	Covariance	,002	-,007	,002
	N	39	39	39

**. Correlation is significant at the 0.01 level (2-tailed).

In Table 15 and Table 16 a significant positive correlation can be found (based on TRQ measure) between C and C_V as well as C and C_C within both test sets. There is a negative correlation between both corpus segments C_V and C_C though not significant. The statistical quality of the built clusters C_V and C_C is again less significant than within the test set CW_{1k}. It can be concluded that the statistical difference between the corpus clusters is lower the smaller the number of terms within the test sets is.

4.3.1.1.4 Excurse: Predictability of the TRQ Plot

Two perspectives can be applied on corpus measure time series: From backward to the past, on the one hand, and from present to future, on the other. The backward perspective is applied when significant thresholds for filtering are used (see Chapter 3.5.4). In this chapter the focus is on predictability of the TRQ time series. For this, the qualities of the TRQ time series must be analysed. The basis data set here was the (not vertically segmented) corpus CW_{5k}. The analysis shown here was originally taken from [Kall05].

A time series analysis resulted in a present auto-correlation of the TRQ plot. For this, the Box-Ljung Test was applied 0:

[13] $$Q_{BL} := N(N+2)\sum_{\tau=1}^{m}\frac{\hat{\rho}_\tau^2}{N-\tau}$$

In 0 is *N* = number of test set elements; ρ_τ = probabilistic Autocorrelation at the distance τ . This is the precondition of the application of an ARIMA model. An ARIMA (1,0,0) model was used with ln conversion of the TRQ.

Fig. 34 shows significant auto correlations for Lag-Numbers<6 and >16. The ACF value first falls down (for Lag-Numbers >10 below zero), and then it rises again.

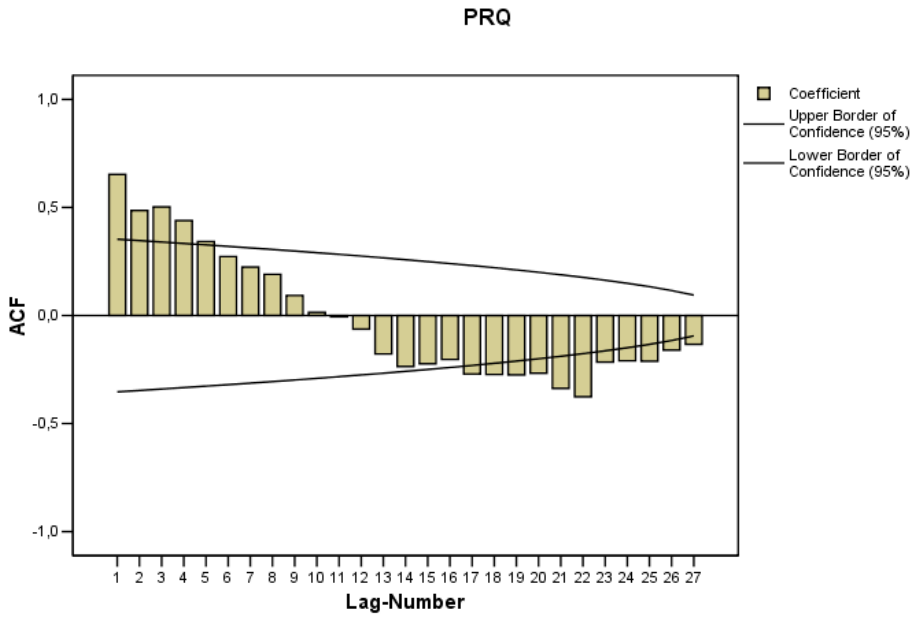


Fig. 34: ACF of TRQ plot

The partial ACF (see Fig. 35) shows that the auto correlation is of 1st order.

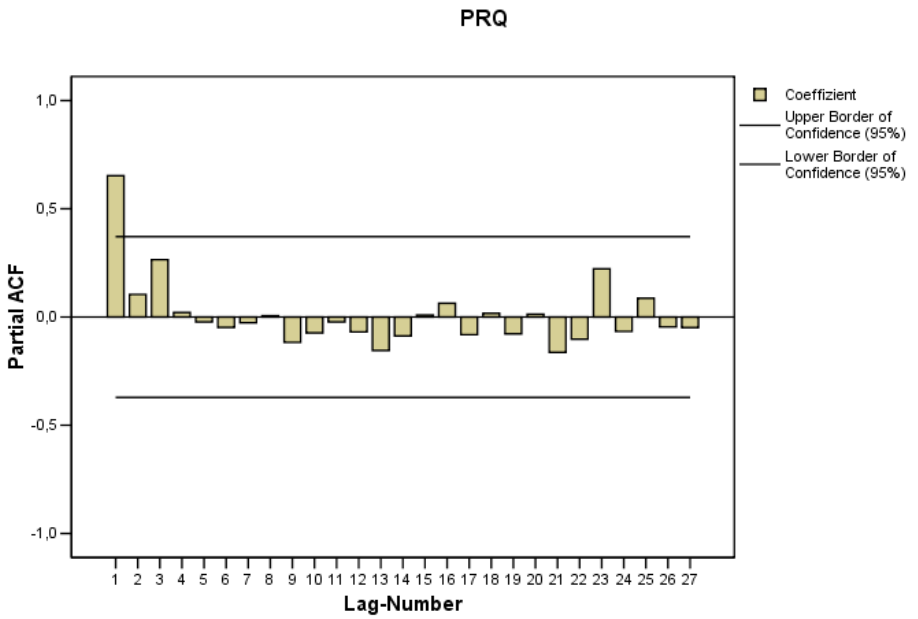


Fig. 35: Partial ACF of TRQ plot

The ACF/PACF quality found mainly fits the condition for an ARIMA (1,0,0) process, where the PACF gives the secure for the correct decision for the apply ability.

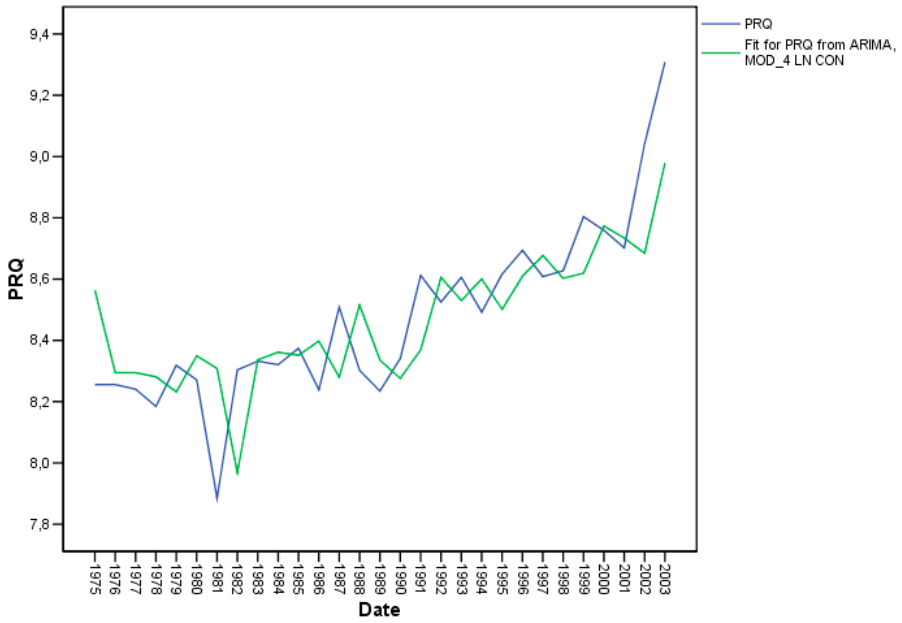


Fig. 36: Graph of TRQ and Fit for TRQ from ARIMA (1,0,0) model with ln Transformation

With the parameters found the TRQ process can be described with Formula 0:

$$\begin{aligned} X_t &= \phi_1 X_{t-1} + E_t \\ &= \phi_1(\phi_1 X_{t-2} + E_{t-1}) + E_t \\ &= \phi_1^2 X_{t-2} + \phi_1 E_{t-1} + E_t \end{aligned} \quad [14]$$

[15] $X_t = 2{,}15 + 0{,}872^2\, X_{t-2} + 0{,}872\, E_{t-1} + E_t$

The residual variance=0,00045 and co-variance=0.018. The Pearson correlation between TRQ and Fit for TRQ from the model is 0.773 at a confidence level of 0.01 (two-sided), which points to a strong relationship. With Formula 0 the TRQ of this specific example can be predicted for future time periods.

4.3.1.2 Statistical analysis of type n corpora summary

Applied to data sets with a high intensity of pre-processing, typical statistical equalities were found that are significant among all four data sets. The correlations of TRQ of each single corpus segment are shown in Table 17.

Table 17: Statistical analysis summary corpus type n

Test Set	Correlation of corpus segment pair C-C_C	Correlation of corpus segment pair C-C_V	Correlation of corpus segment pair C_C-C_V
CW_{5k}	Negative, middle, significant	Positive, strong, significant	Negative, strong, significant
CW_{1k}	~ 0	Positive, middle, significant	Negative, middle, significant
$Al1k_{S1}$	Positive, middle, significant	Positive, strong, significant	Negative, weak
$Al1k_{S2}$	Positive, middle, significant	Positive, strong, significant	Negative, weak

4.3.1.3 Distribution Analysis of applied Taxonomies on type n corpora

As discussed in Chapter 3.6.1 *f.*, taxonomies have a pre-selection effect on the data, especially in the case where they do not cover each term. Measuring in this context may focus here on two different perspectives: One could be the quality of the taxonomy itself; another could be the internal corpus structure and their semantic evolution. The perspective choice is due to the aim of the analysis. Unfortunately none of these perspectives may occur purely in reality, so that in most cases a mixture of both determining factors will be present. A statistically neutral approach is chosen here, which permits covering both perspectives, even by declaring one or both (the taxonomy or the corpus) as fixed. In Table 52 (see Appendix) the number of matching terms, when using different types of taxonomies, can be seen as “Dim” as an example of a general approach with many matching terms.

“Dim_Mertens_alone” here is an example of ex-ante definition of only a very limited number of terms observed. The application of both taxonomies on the same test sets leads to very contrary results regarding the volatility in the number of matched terms over the corpus time segments. Regardless of whether applied on CW_{5k} or CW_{1k} test set, taxonomy “Dim” always covers a very even number of terms per year. The pre-selected terms in “Dim_Mertens_alone” lead to a very volatile assignment.

For the Al1k test set (see Table 53 in Appendix) the “Dim” taxonomy with domain-specific concepts was used and the “CC_Dim_alone” taxonomy with persistent concepts. As was to be expected, these taxonomies evenly cover

terms throughout the timeline. Only for “Dim” the number of assigned terms in corpus segment C_C was very volatile between values “0” to “15” due to the limited size of the AI1k test sets.

4.3.1.3.1 Distribution analysis of CW corpus test set CW_{5k}

In this chapter the statistical qualities of the resulting corpus data sets are analysed when different taxonomies are applied.

Table 54 (see Appendix) documents that the “No suffix” taxonomy combination (see first case in Table 9), which represents the optimal pre-processed corpus, allows an assignment of terms with the lowest absolute and percentage value of range and standard deviation.

The taxonomies used here also led to smoothing of corpus segment differences. Where C and C_V clearly have different statistical characteristics (see Table 48), after application of the taxonomies the differences between both segments are completely eliminated (see

Table 54). This extreme filtering capability results from the filtering character of the used taxonomies. As described in

Table 4, only terms were recognized that appear at least 21 times in a certain year. An interesting fact here is that the application of the taxonomies and the exact computing of the constant terms (and their filtering from C) led to equal results.

Especially with the “Dim_Mertens” taxonomy[32] it can be seen that a very focused taxonomy leads to low matching rates if applied to the large CW_{5k} corpus. The high volatility in term coverage can be comprehended in Table 52.

Even without any knowledge of the issue date of Mertens’ analysis, a strong peak in assigned terms in the early 1990s points to the main emphasis of his domain perspective that is materialized in the choice of the observed concepts.

Contrary the “Dim” taxonomy in the opposite represents a very evenly distributed coverage of terms, especially when observing the C_C segment in Table 52, which mainly contains terms that belong either to the language itself or to the domain. The CV segment shows a typical graph with an “S”-like shape. The “Dim” dimension, which was constructed by assigning all terms occurring more than 20 times within the corpus, covers more terms in the first periods observed. After covering fewer terms an increase in terms assigned follows before the rate again decreases. Remembering the TRQ plot for this test sets (see Fig. 36) raises the question regarding the statistical dependency between the TRQ graph and the graph of assigned types by the taxonomies. This is analysed with the application of a correlation analysis with significance tests (see

[32] The Dim_Mertens taxonomy consists of only 10 terms.

Table 18). With the reference test set CW_{5k} a significant strong negative correlation of -0,693 between the two observed time series at 99% level of significance (both-sided) was found .Based on these results the conclusion that can be made is that the fewer different terms that occur within a corpus the higher the number of terms is that are assigned to an applied taxonomy, and vice versa. With a higher absolute value |-0,803| than for the whole corpus this negative correlation was also found for segment C_V. This shows that a filtering of C_C corpus segment from C raises the negative correlation in the observed test set CW_{5k}.

Table 18: Analysis of correlation between types and assigned terms per year by taxonomy Dim of CW_{5k} corpus

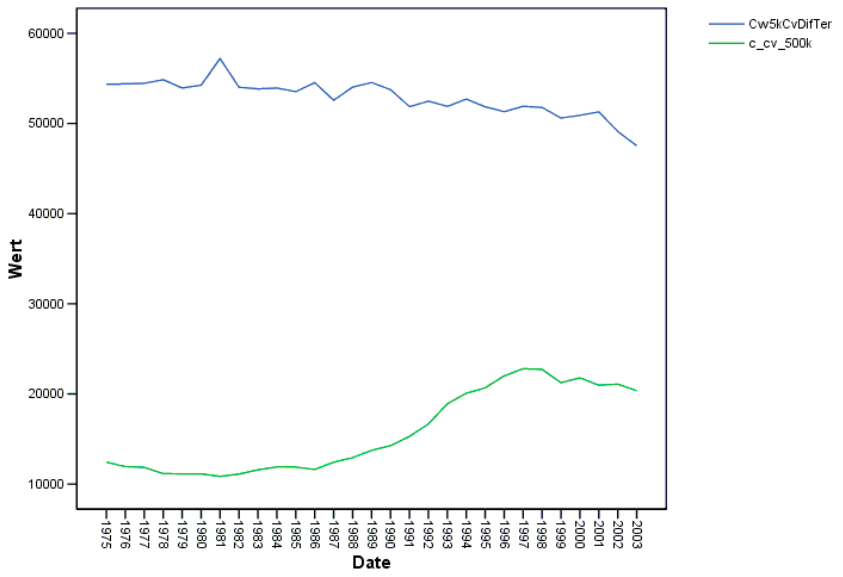
Segment	Results of correlation analysis
C	

Korrelationen

		Cw5kCDifTer	c_ges_500k
Cw5kCDifTer	Korrelation nach Pearson	1	-,693**
	Signifikanz (2-seitig)		,000
	Quadratsummen und Kreuzprodukte	106368905,0	-159178279
	Kovarianz	3798889,466	-5684938,5
	N	29	29
c_ges_500k	Korrelation nach Pearson	-,693**	1
	Signifikanz (2-seitig)	,000	
	Quadratsummen und Kreuzprodukte	-159178279	496558057
	Kovarianz	-5684938,521	17734216,3
	N	29	29

**. Die Korrelation ist auf dem Niveau von 0,01 (2-seitig) signifikant.

C_V



Korrelationen

		Cw5kCvDifTer	c_cv_500k
Cw5kCvDifTer	Korrelation nach Pearson	1	-,803**
	Signifikanz (2-seitig)		,000
	Quadratsummen und Kreuzprodukte	106368905,0	-2,0E+08
	Kovarianz	3798889,466	-7088484
	N	29	29
c_cv_500k	Korrelation nach Pearson	-,803**	1
	Signifikanz (2-seitig)	,000	
	Quadratsummen und Kreuzprodukte	-198477543,6	5,74E+08
	Kovarianz	-7088483,698	20501941
	N	29	29

**. Die Korrelation ist auf dem Niveau von 0,01 (2-seitig) signifikant.

4.3.1.3.2 Distribution analysis of CW corpus test set CW_{1k}

In this chapter the statistical qualities of the resulting corpus data sets are analysed when different taxonomies are applied on CW_{1k} test set.

With comparable results to the analysis of test set CW_{5k} Table 55 (see Appendix) documents that the “No suffix” taxonomy combination (see first case in Table 9), which represents the optimal pre-processed corpus, allows an assignment of terms with the lowest absolute and percentage value of range and standard deviation.

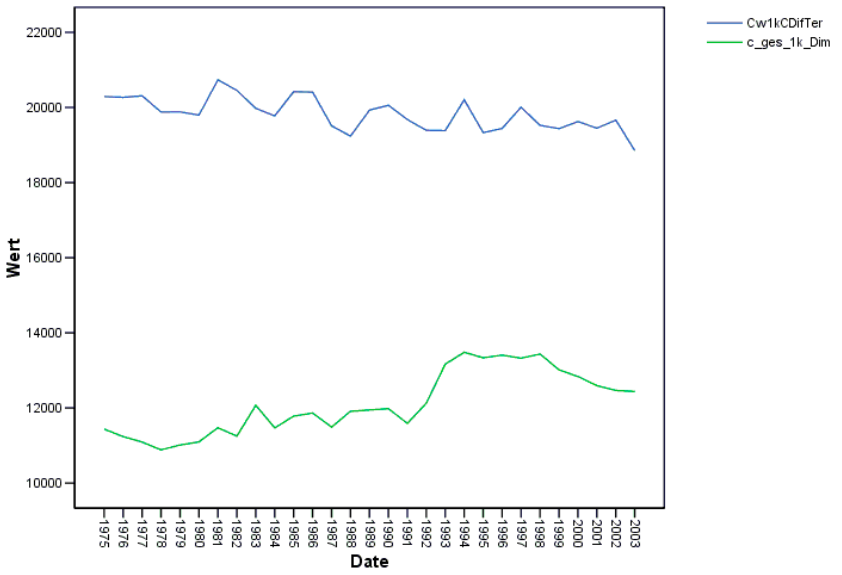
The taxonomies used here also led to smoothing of corpus segment differences. Where C and C_V clearly have different statistical characteristics (see

Table 49), after application of the taxonomies the difference between both these segments are completely eliminated (see Table 55 in Appendix). This extreme filtering capability results from the filtering character of the taxonomies used. As described in

Table 4, only terms were recognized that appeared at least 21 times within a certain year. An interesting fact here is that the application of the taxonomies and the exact computing of the constant terms (and their filtering from C) led to equal results.

The “Dim_Mertens” taxonomy leads again to low matching rates of the CW_{1k} corpus with higher volatility in term assignment than with CW_{5k} (see Table 52). The strong peak in assigned terms in the early 1990s found for CW_{5k} is also present for CW_{1k}. The “Dim” taxonomy has a quite evenly distributed coverage of terms, especially when observing the C_C segment in Table 52. Compared to the reference test set CW_{5k} a significant strong negative correlation was found of -0.451 between the two observed time series at 95% level of significance (both-sided). This result was weaker than that derived from CW_{5k}. The absolute value of negative correlation was also higher for segment C_V.

Table 19: Analysis of correlation between types and assigned terms per year by taxonomy Dim of CW_{1k} corpus

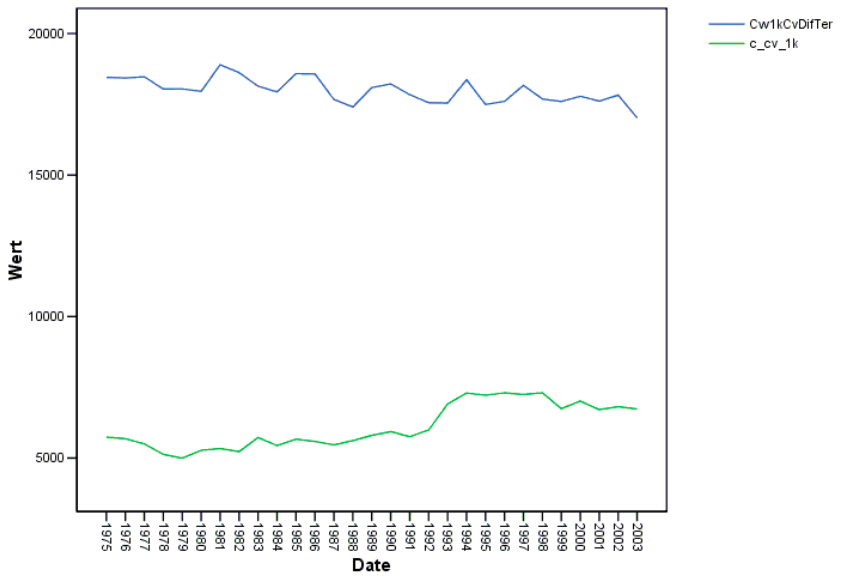
Segment	Results of correlation analysis
C	

Korrelationen

		Cw1kCDifTer	c_ges_ 1k_Dim
Cw1kCDifTer	Korrelation nach Pearson	1	-,451*
	Signifikanz (2-seitig)		,014
	Quadratsummen und Kreuzprodukte	5468865,034	-4680475
	Kovarianz	195316,608	-167160
	N	29	29
c_ges_1k_Dim	Korrelation nach Pearson	-,451*	1
	Signifikanz (2-seitig)	,014	
	Quadratsummen und Kreuzprodukte	-4680475,483	2,0E+07
	Kovarianz	-167159,839	703402,2
	N	29	29

*. Die Korrelation ist auf dem Niveau von 0,05 (2-seitig) signifikant.

C_V



Korrelationen

		Cw1kCvDifTer	c_cv_1k
Cw1kCvDifTer	Korrelation nach Pearson	1	-,471**
	Signifikanz (2-seitig)		,010
	Quadratsummen und Kreuzprodukte	5468865,034	-4523706
	Kovarianz	195316,608	-161561
	N	29	29
c_cv_1k	Korrelation nach Pearson	-,471**	1
	Signifikanz (2-seitig)	,010	
	Quadratsummen und Kreuzprodukte	-4523705,552	1,7E+07
	Kovarianz	-161560,913	602119,8
	N	29	29

**. Die Korrelation ist auf dem Niveau von 0,01 (2-seitig) signifikant.

4.3.1.3.3 Distribution analysis of Allianz corpus test sets $Al1k_{S1}$ and $Al1k_{S2}$

In this chapter the statistical qualities of the resulting corpus data sets are analysed when different taxonomies are applied on Al1k test set.

Whereas the quantitative relations of assigned terms comparable to the CW test sets, it is remarkable for these two small-size test sets that the absolute number of assigned terms in C_V is very close to the number of assigned terms within the whole corpus (see Table 56 in Appendix). These test sets of type “n” of a small number of yearly terms are nearly completely assigned to the domain-related taxonomy “Dim”.

The applied analysis documented in Table 20 shows comparable results in correlation between the number of different terms and the number of assigned terms to a taxonomy compared with CW test sets within the whole corpus.

One main difference was found in correlation for the C_V segment. A very weak negative correlation was found in both test sets (see Table 20 and Table 21). The factor responsible for this is probably the limited size of the AI1k corpus.

Table 20: Analysis of correlation between types and assigned terms per year by taxonomy Dim of $AI1k_{S1}$ corpus

Segment	Results of correlation analysis
C	

Korrelationen

		AI100S1CDifTer	c_ges_all_1k1
AI100S1CDifTer	Korrelation nach Pearson	1	-,589**
	Signifikanz (2-seitig)		,000
	Quadratsummen und Kreuzprodukte	12952,000	-9598,000
	Kovarianz	417,806	-309,613
	N	32	32
c_ges_all_1k1	Korrelation nach Pearson	-,589**	1
	Signifikanz (2-seitig)	,000	
	Quadratsummen und Kreuzprodukte	-9598,000	20507,969
	Kovarianz	-309,613	661,547
	N	32	32

**. Die Korrelation ist auf dem Niveau von 0,01 (2-seitig) signifikant.

Cv

Korrelationen

		Al100S1Cv DifTer	c_cv_all_1k1
Al100S1CvDifTer	Korrelation nach Pearson	1	-,083
	Signifikanz (2-seitig)		,652
	Quadratsummen und Kreuzprodukte	8092,469	-1073,844
	Kovarianz	261,047	-34,640
	N	32	32
c_cv_all_1k1	Korrelation nach Pearson	-,083	1
	Signifikanz (2-seitig)	,652	
	Quadratsummen und Kreuzprodukte	-1073,844	20704,719
	Kovarianz	-34,640	667,894
	N	32	32

Table 21: Analysis of correlation between types and assigned terms per year by taxonomy Dim of $AI1k_{S2}$ corpus

Segment	Results of correlation analysis
C	

Korrelationen

		AI100S2CDifTer	c_ges_all_1k2
AI100S2CDifTer	Korrelation nach Pearson	1	-,576**
	Signifikanz (2-seitig)		,001
	Quadratsummen und Kreuzprodukte	13689,875	-12643,3
	Kovarianz	441,609	-407,849
	N	32	32
c_ges_all_1k2	Korrelation nach Pearson	-,576**	1
	Signifikanz (2-seitig)	,001	
	Quadratsummen und Kreuzprodukte	-12643,313	35236,719
	Kovarianz	-407,849	1136,668
	N	32	32

**. Die Korrelation ist auf dem Niveau von 0,01 (2-seitig) signifikant.

Cv

Korrelationen

		Al100S2Cv DifTer	c_cv_all_1k2
Al100S2CvDifTer	Korrelation nach Pearson	1	-,079
	Signifikanz (2-seitig)		,666
	Quadratsummen und Kreuzprodukte	11738,969	-1556,406
	Kovarianz	378,676	-50,207
	N	32	32
c_cv_all_1k2	Korrelation nach Pearson	-,079	1
	Signifikanz (2-seitig)	,666	
	Quadratsummen und Kreuzprodukte	-1556,406	32745,719
	Kovarianz	-50,207	1056,314
	N	32	32

4.3.1.4 Distribution analysis of applied taxonomies on type n corpora summary

A negative correlation between the number of different terms and the number of terms assigned to a given taxonomy was found for each yearly segment (see Table 22).

Table 22: Correlation between types and assigned terms per year by taxonomy for corpus type n

Test Set	Taxonomy	Corpus segment C	Corpus segment C_V
CW_{5k}	Dim	Negative, middle, significant	Negative, strong, significant
CW_{1k}	Dim	Negative, middle, significant	Negative, middle, significant
$Al1k_{S1}$	Dim	Negative, middle, significant	Negative, weak
$Al1k_{S2}$	Dim	Negative, middle, significant	Negative, weak

This correlation was always middle and significant for all C segments within all test sets. For segment C_V a negative correlation was found, but with falling absolute value and falling statistical support the smaller the size of test set was that were tested.

This dependency can be interpreted as follows: The lower the number of different terms within a corpus, the higher the absolute number of assigned terms to a given taxonomy, and vice versa.

4.3.1.5 Semantic analysis of type n corpora

In the previous chapters statistical qualities were analysed that can act as indicators for the intensity of pre-processing of certain test sets. In this chapter the extraction of knowledge will be done according to the methods described in Chapter 3.2.3.2. The domain knowledge introduced in Chapters 3.1.3 and 3.1.4 is not used as a strict benchmark, but as a basis for an evaluation of results from a domain-expert perspective. The following chapters focus on each test set separately. A summary is given in Chapter 4.3.1.6.

4.3.1.5.1 Semantic analysis of CW corpus test set CW_{5k}

In the following diagrams the progress paths of certain concepts are compared directly to the ones introduced by the work of Mertens [Mert95]. It must

be considered that the time period of the CW_{5k} graph covers a range from 1975 to 2003, not only 1975 to 1994 as in [Mert95]. No statistical-proved comparison is given here between Mertens' results and the results derived from the approach introduced, but Mertens' results were used for a rough plausibility check, whether the results match in general or if there are extreme differences present. The graphs for the occurrence of each concept used by Mertens are shown in the following figures. Although he used a "share of mentioned topics" measure, the results are comparable because the CountSum measure used in the CW_{5k} is equally defined due to the constants of number of terms in yearly corpus segments.

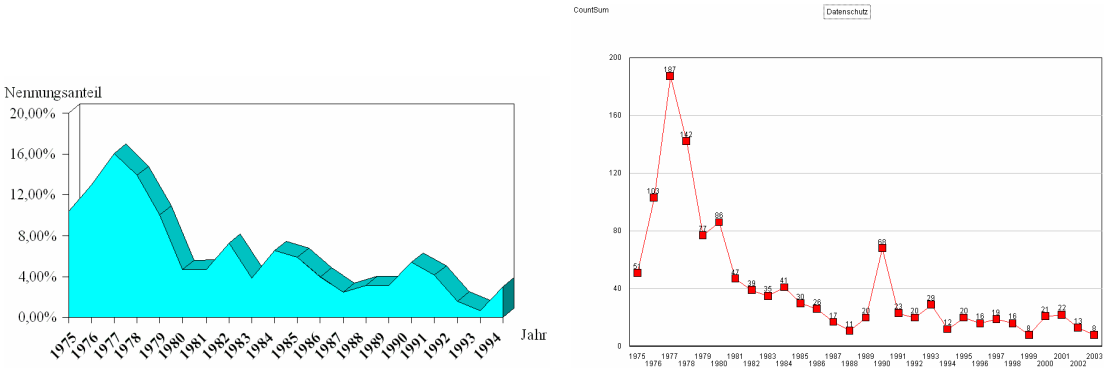


Fig. 37: The progress of the concept "Datenschutz" from [Mert95] and extracted from CW_{5k} corpus

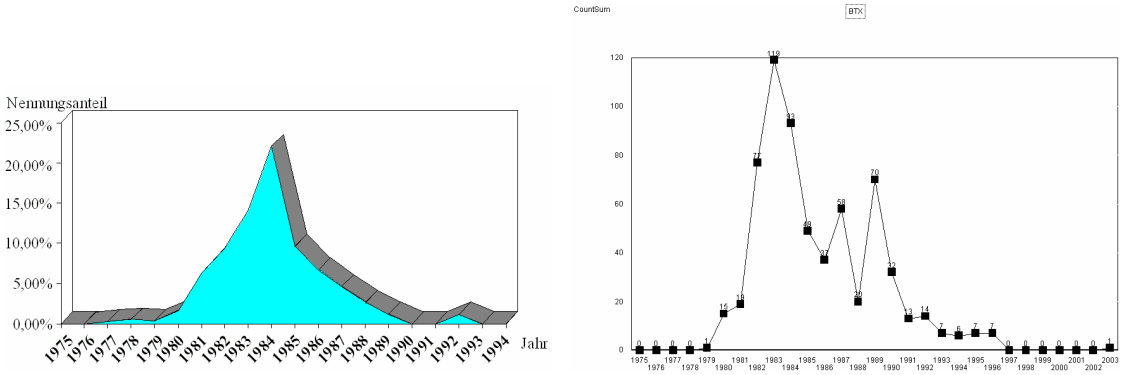


Fig. 38: The progress of the concept "BTX" from [Mert95] and extracted from CW_{5k} corpus

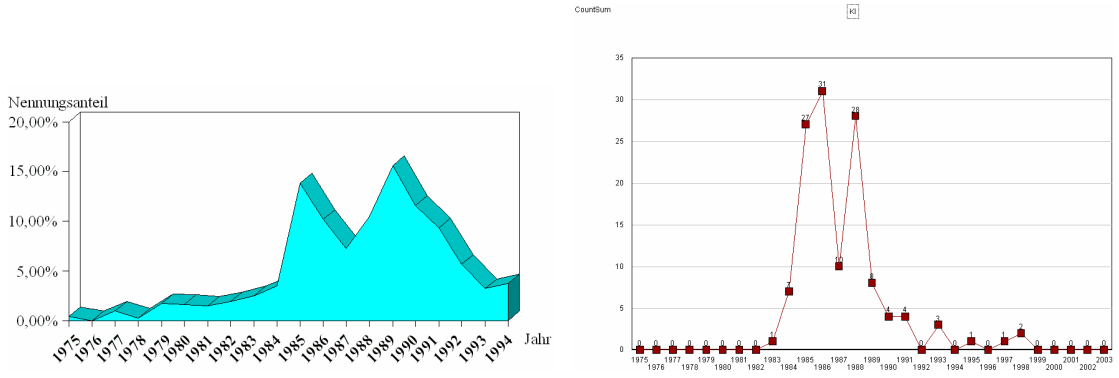


Fig. 39: The progress of the concept "KI" from [Mert95] and extracted from CW_{5k} corpus

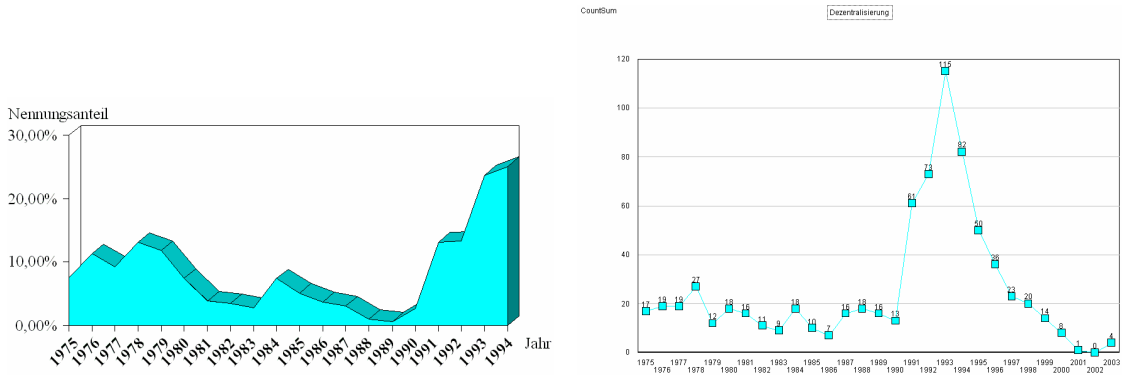


Fig. 40: The progress of the concept "Dezentralisierung" from [Mert95] and extracted from CW_{5k} corpus

The progress paths for the concepts “Datenschutz”, “BTX”, ”KI” and “Dezentralisierung” automatically derived from the CW_{5k} corpus are very similar to those found by Mertens with manual processing. Differences were present for the other concepts:

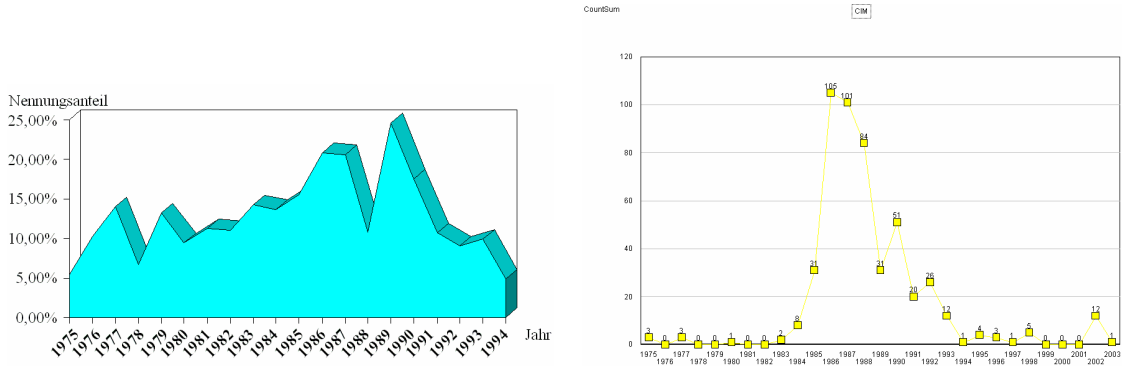


Fig. 41: The progress of the concept "CIM" from [Mert95] and extracted from CW_{5k} corpus

“CIM” was recognized as a concept with a high persistence during all periods, but found to be hype in the mid-1980s with the automatic approach.

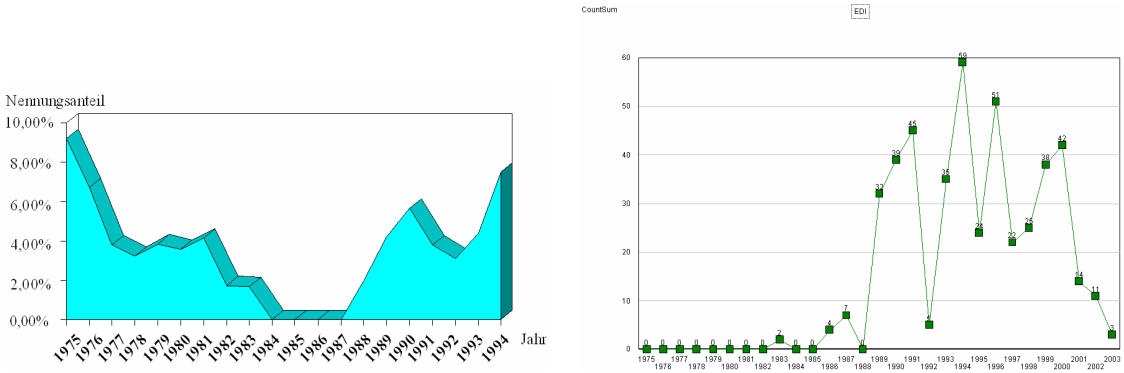


Fig. 42: The progress of the concept "EDI" from [Mert95] and extracted from CW_{5k} corpus

The concepts “EDI” and “Outsourcing” were mentioned by Mertens as important topics in the early 1970s, but not found to be dominating until the mid-1980s with the automatic method.

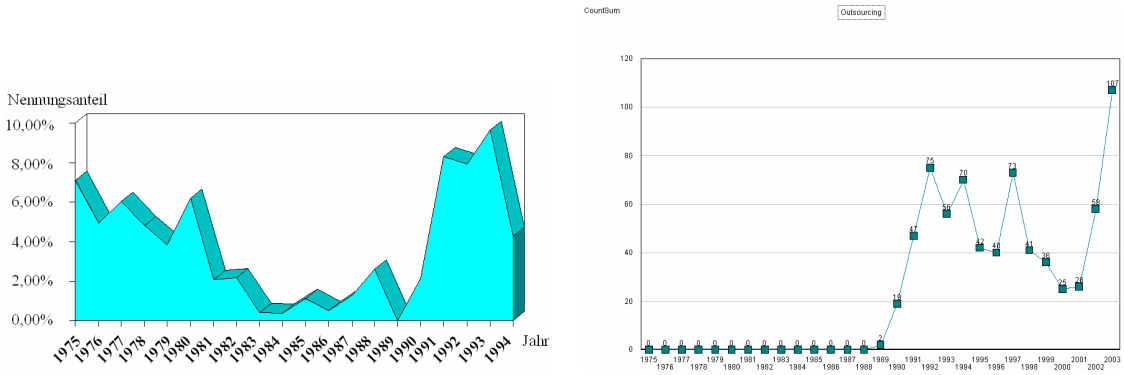


Fig. 43: The progress of the concept "Outsourcing" from [Mert95] and extracted from CW_{5k} corpus

It must be considered here that only a pre-selection of terms or concepts based on expert know-how led to the taxonomies observed in “Dim_Mertens”. This approach is biased by definition with individual knowledge and expectation. It is a kind of retrospective view of progress paths of concepts that were ex-ante defined. It can be summarized that even if similar concepts were extracted as leading during certain periods, a divergence was in some cases present. This is to be attributed to the fact that in [Mert95] the

extraction was done using the predominant factor of expert knowledge, whereas the CW_{5k} was analysed by a purely quantitative approach.

Which domain knowledge did the segmenting approach based on TRQ threshold (see Chapter 4.2.2) extract? In general, the approach permits the extraction of significant aggregated concepts from a text collection with the opportunity to drill down to term level. Previous knowledge, in contrast with expert-based approaches, is not used.

Table 23 documents the first ten leading aggregated concepts within corpus segments and drill down to term level in CW_{5k}. This is a limited list, only for an introduction of results. For a complete list refer to Appendix (see Table 71). The leading concepts are shown separately for each corpus segment C_C and C_V[33]. As an example of a non-persistent term the occurrence period of the term “Chipcom” is shown with its period of occurrence from 1987-1998. Another drill-down is done for the concept “Vendor”.

Table 23: CW_{5k}, first ten leading aggregated the concepts within corpus segments and drill down to term level

	CW_{5k}		
Date	1975	1988	2003
Cc_CountThresU_Dim	Currency	Currency	Currency
	Vendor	Vendor	Economy
	Profession	Economy	Vendor
	IT	IT	Geography
	Geography	Customer	IT
	Economy	Business	Business
	OS	Geography	Performance
	Customer	Profession	Profession
	Business	Performance	Science
	Performance	Science	Customer
Cv_CountThresU_Dim	Currency	OS	Currency
	Vendor	Norm	OS
	IT	Institute	ProgLanguage
	ITProduct	Vendor	Vendor
	OS	ITProduct	Profession
	ProgLanguage	ProgLanguage	Institute
	Performance	IT	IT
	Institute	Currency	ITProduct
	Profession	Event	Economy
	Name	Performance	Performance
Chipcom.TermFirstOcc	1987		
Chipcom.TermLastOcc	1998		
Cc_CountThresU_Vendor	IBM	IBM	IBM
	Siemens	DEC	HP
	Nixdorf	Siemens	Intel
	Bull	IBMs	Siemens
	Philips	Digital	Hewlett-Packard
	Xerox	HP	
	Digital	Hewlett-Packard	
	NCR	Bull	
		Nixdorf	

[33] Even if the names of concepts are equal within C_C and C_V, the terms assigned to these concepts are not the same, because the terms were automatically assigned to one of both corpus segments due to their persistence in time.

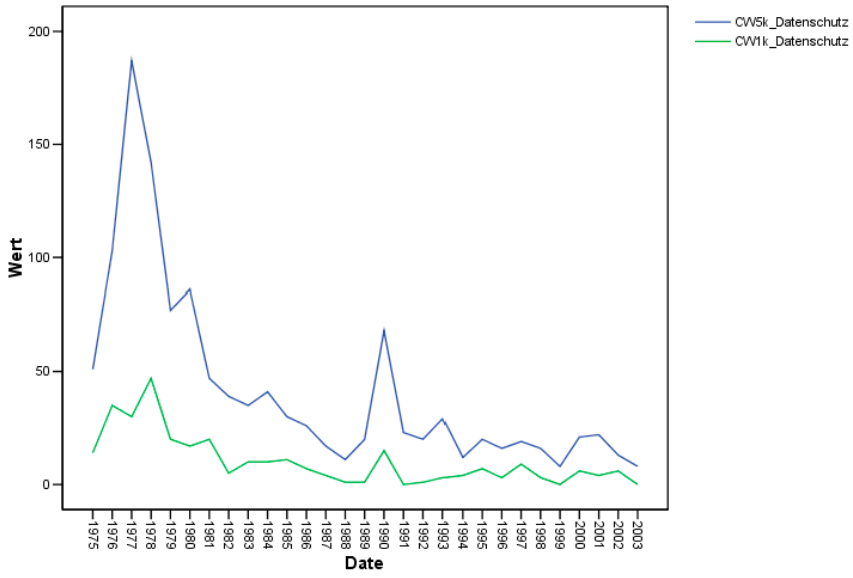
	CW_{5k}		
Date	1975	1988	2003
		NCR	
Cv_CountThresU_Vendor	Honeywell	Sun	Microsoft
	Univac	Apple	SAP
	Unidata	Microsoft	Sun
	Kienzle	Bundespost	Oracle
	Bundespost	Novell	SCO
	CDC	Unisys	Abb
	Burroughs	Oracle	Microsofts
	Singer	Wang	Telekom
	Sperry	Apollo	Suse
	Interdata	Amdahl	EDS

In the corpus segments C_C and C_V quite similar concepts can be found as dominating the first ranks (e.g., “Currency, “Vendor”, “Norm”, “OS”, and “Economy”). The concept “OS” (operating system) is an example of a very volatile concept in CW_{5k}, which plays no role in C_C but leads during several periods in C_V. This fact indicates the high volatility in operating systems market positions within the last 30 years.

In later chapters the domain knowledge extracted from test sets with different sizes or different intensity of pre-processing is compared to the reference results shown in Table 23.

4.3.1.5.2 Semantic analysis of CW corpus test set CW_{1k}

Compared to CW_{5k}: How similar is the extracted knowledge from the used test sets? With “Dim_Mertens” a taxonomy is available that focuses on a limited number of concepts. The occurrence of these concepts is measurable within both test sets. As a simple indicator of similarity Pearson’s correlation is used here. It is expected to find high and significant correlation factors if both test sets are similar. Fig. 44 to Fig. 50 show graphs of the correlations and computed values.

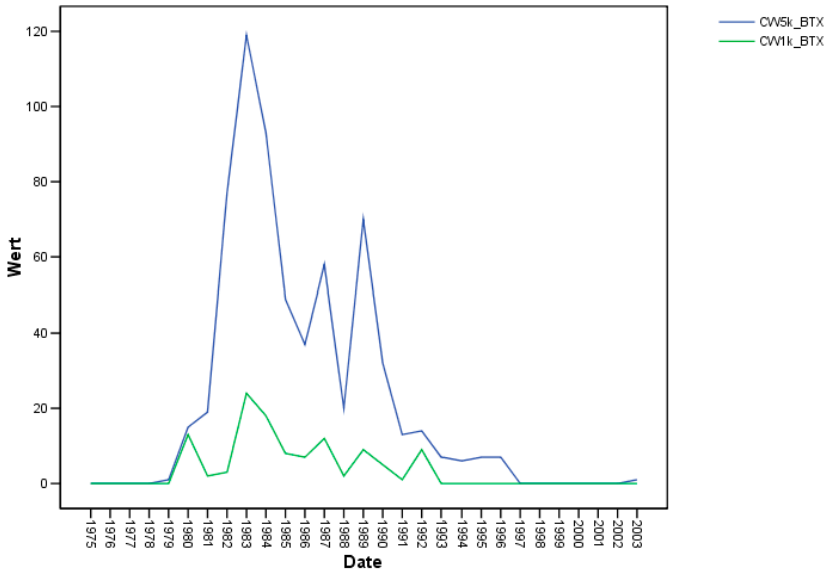


Korrelationen

		CW5k_ Datenschutz	CW1k_ Datenschutz
CW5k_Datenschutz	Korrelation nach Pearson	1	,888**
	Signifikanz (2-seitig)		,000
	Quadratsummen und Kreuzprodukte	49486,828	11790,138
	Kovarianz	1767,387	421,076
	N	29	29
CW1k_Datenschutz	Korrelation nach Pearson	,888**	1
	Signifikanz (2-seitig)	,000	
	Quadratsummen und Kreuzprodukte	11790,138	3558,690
	Kovarianz	421,076	127,096
	N	29	29

**. Die Korrelation ist auf dem Niveau von 0,01 (2-seitig) signifikant.

Fig. 44: Correlation between progresses of the concept "Datenschutz" extracted from CW_{5k} and from CW_{1k} corpus

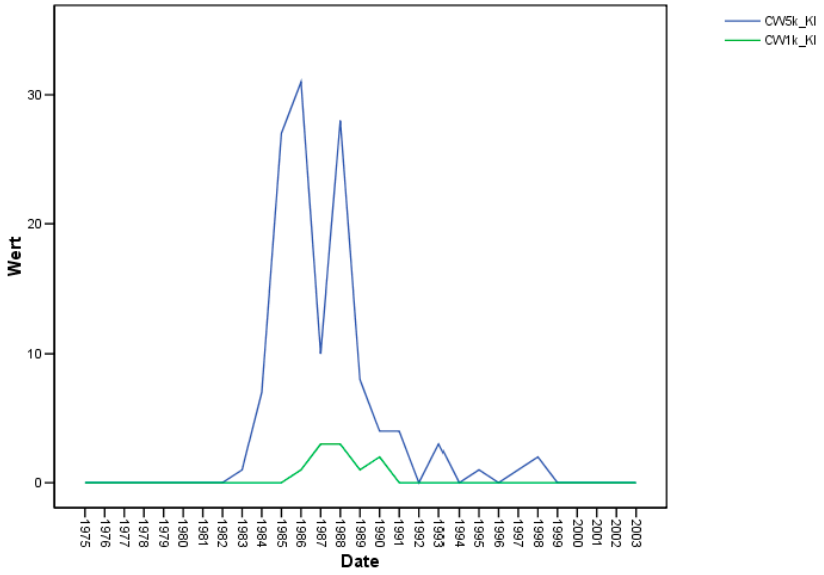


Korrelationen

		CW5k_BTX	CW1k_BTX
CW5k_BTX	Korrelation nach Pearson	1	,853**
	Signifikanz (2-seitig)		,000
	Quadratsummen und Kreuzprodukte	28987,310	4796,724
	Kovarianz	1035,261	171,312
	N	29	29
CW1k_BTX	Korrelation nach Pearson	,853**	1
	Signifikanz (2-seitig)	,000	
	Quadratsummen und Kreuzprodukte	4796,724	1090,690
	Kovarianz	171,312	38,953
	N	29	29

**. Die Korrelation ist auf dem Niveau von 0,01 (2-seitig) signifikant.

Fig. 45: Correlation between progresses of the concept "BTX" extracted from CW_{5k} and from CW_{1k} corpus

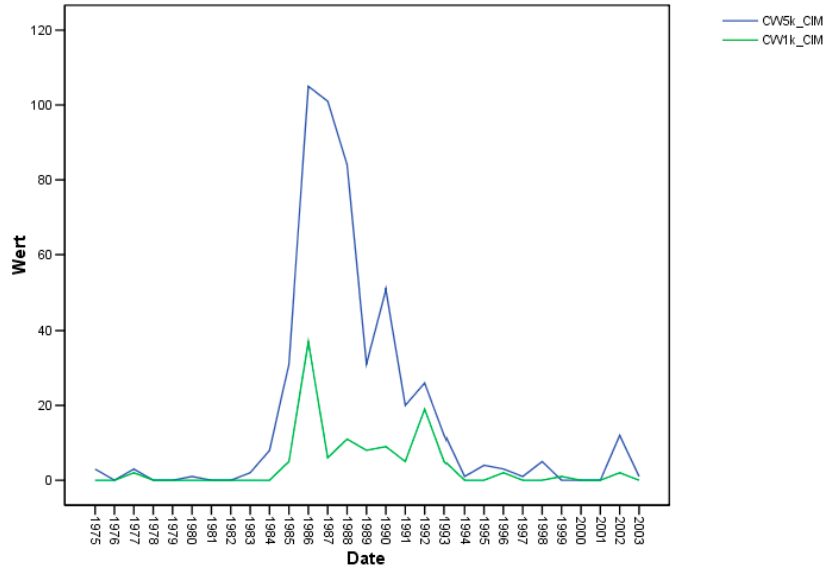


Korrelationen

		CW5k_KI	CW1k_KI
CW5k_KI	Korrelation nach Pearson	1	,554**
	Signifikanz (2-seitig)		,002
	Quadratsummen und Kreuzprodukte	2178,828	117,207
	Kovarianz	77,815	4,186
	N	29	29
CW1k_KI	Korrelation nach Pearson	,554**	1
	Signifikanz (2-seitig)	,002	
	Quadratsummen und Kreuzprodukte	117,207	20,552
	Kovarianz	4,186	,734
	N	29	29

**. Die Korrelation ist auf dem Niveau von 0,01 (2-seitig) signifikant.

Fig. 46: Correlation between progresses of the concept "KI" extracted from CW_{5k} and from CW_{1k} corpus

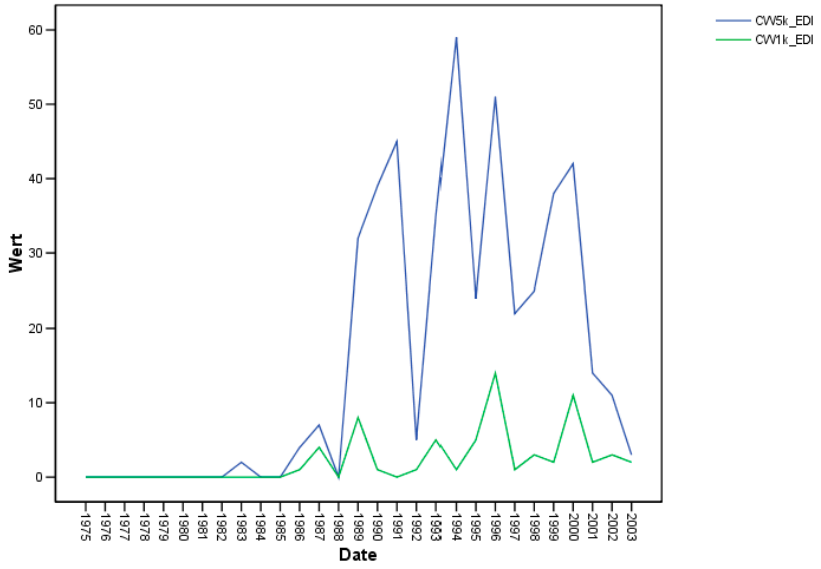


Korrelationen

		CW5k_CIM	CW1k_CIM
CW5k_CIM	Korrelation nach Pearson	1	,765**
	Signifikanz (2-seitig)		,000
	Quadratsummen und Kreuzprodukte	25515,034	5016,655
	Kovarianz	911,251	179,166
	N	29	29
CW1k_CIM	Korrelation nach Pearson	,765**	1
	Signifikanz (2-seitig)	,000	
	Quadratsummen und Kreuzprodukte	5016,655	1687,448
	Kovarianz	179,166	60,266
	N	29	29

**. Die Korrelation ist auf dem Niveau von 0,01 (2-seitig) signifikant.

Fig. 47: Correlation between progresses of the concept "CIM" extracted from CW_{5k} and from CW_{1k} corpus

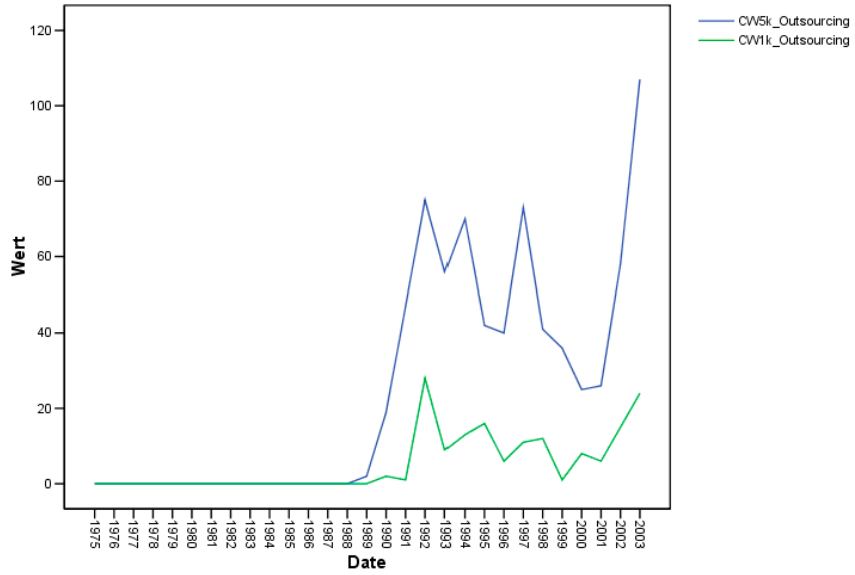


Korrelationen

		CW5k_EDI	CW1k_EDI
CW5k_EDI	Korrelation nach Pearson	1	,592**
	Signifikanz (2-seitig)		,001
	Quadratsummen und Kreuzprodukte	9956,759	1091,241
	Kovarianz	355,599	38,973
	N	29	29
CW1k_EDI	Korrelation nach Pearson	,592**	1
	Signifikanz (2-seitig)	,001	
	Quadratsummen und Kreuzprodukte	1091,241	340,759
	Kovarianz	38,973	12,170
	N	29	29

**. Die Korrelation ist auf dem Niveau von 0,01 (2-seitig) signifikant.

Fig. 48: Correlation between progresses of the concept "EDI" extracted from CW_{5k} and from CW_{1k} corpus



Korrelationen

		CW5k_ Outsourcing	CW1k_ Outsourcing
CW5k_Outsourcing	Korrelation nach Pearson	1	,879**
	Signifikanz (2-seitig)		,000
	Quadratsummen und Kreuzprodukte	26291,793	5877,931
	Kovarianz	938,993	209,926
	N	29	29
CW1k_Outsourcing	Korrelation nach Pearson	,879**	1
	Signifikanz (2-seitig)	,000	
	Quadratsummen und Kreuzprodukte	5877,931	1701,310
	Kovarianz	209,926	60,761
	N	29	29

**. Die Korrelation ist auf dem Niveau von 0,01 (2-seitig) signifikant.

Fig. 49: Correlation between progresses of the concept "Outsourcing" extracted from CW_{5k} and from CW_{1k} corpus

Korrelationen

		CW5k_ Dezentrali sierung	CW1k_ Dezentrali sierung
CW5k_Dezentralisierung	Korrelation nach Pearson	1	,711**
	Signifikanz (2-seitig)		,000
	Quadratsummen und Kreuzprodukte	2301,310	491,103
	Kovarianz	82,190	17,539
	N	29	29
CW1k_Dezentralisierung	Korrelation nach Pearson	,711**	1
	Signifikanz (2-seitig)	,000	
	Quadratsummen und Kreuzprodukte	491,103	207,034
	Kovarianz	17,539	7,394
	N	29	29

**. Die Korrelation ist auf dem Niveau von 0,01 (2-seitig) signifikant.

Fig. 50: Correlation between progresses of the concept "Dezentralisierung" extracted from CW_{5k} and from CW_{1k} corpus

All correlations for all concepts were significant at 99%, but the absolute correlation values for different concepts had a range between less than 0.6 up to 0.888. A strong coupling between both test sets was present, but dependent on certain terms or concepts.

Table 24 documents the semantic analysis of CW_{1K}, with the first ten leading aggregated concepts within corpus segments and drill down to term level. For a complete list refer to Appendix (see Table 72):

Table 24: CW_{1k}, first ten leading aggregated concepts within corpus segments and drill down to term level

	CW_{1k}		
Date	1975	1988	2003
Cc_CountThresU_Dim	Vendor	Vendor	Currency
	Currency	Currency	Economy
	IT	IT	Geography
	Business	Economy	Vendor
	Economy	Business	IT
	Profession	Geography	Business
	Geography	Customer	
	Customer		
Cv_CountThresU_Dim	OS	OS	Currency
	Vendor	Vendor	OS
	Customer	Event	Vendor
	Institute	IT	ProgLanguage
	IT	ITProduct	Institute
	ITProduct	Geography	Profession
	Currency	Norm	Customer
	Economy	Performance	Economy
	Geography	Science	IT
	Name	Business	ITProduct
Chipcom.TermFirstOcc	1989		
Chipcom.TermLastOcc	1996		
Cc_CountThresU_Vendor	IBM	IBM	IBM
	Siemens	Siemens	HP
Cv_CountThresU_Vendor	Unidata	DEC	SAP
	Nixdorf	Sun	Microsoft
	Honeywell	Apple	Sun
	Univac	Intel	Dell
	BASF	Microsoft	Oracle
	Olivetti	Oracle	SCO
	Xerox	Bull	Suse
	Burroughs	Nixdorf	Lexmark
	Kienzle	Fujitsu	Siebel
	Bull	Toshiba	Microsofts

Documented in Table 24 minor differences in extracted concepts were found here compared to the results extracted from CW_{5k} (see Table 23). In general, fewer aggregated concepts were found than past the CountThres threshold (especially in corpus segment C_C) and the results extracted from CW_{5k} are less precise than that. The lower and upper dates for TermFirstOcc and TermLastOcc for term “Chipcom” indicate a range of occurrence from 1987-1998 for CW_{5k} and 1989-1996 for CW_{1k}. It’s possible to conclude therefore that the reduction of corpus size partly led to incomplete and in result, imprecise information. Of course it is possible that results extracted from CW_{5k} are inherent in the absolute truth, but they are closer to it than results extracted from smaller corpora.

4.3.1.5.3 Semantic analysis of Allianz corpus test sets $Al1k_{S1}$ and $Al1k_{S2}$

Table 25 shows the semantic analysis of $Al1k_{S1}$ and $Al1k_{S2}$, with the first ten leading aggregated concepts within corpus segments and drill down to term level (1962-1971). Additionally, two terms were observed: “ELVIA” and “Cornhill”. Both are insurance companies that were acquired by Allianz in 1995 (ELVIA) and 1986 (Cornhill), as mentioned in Fig. 13. The values of TermFirstOcc are expected to be found accordingly for both terms.

Table 25: Al1k, leading aggregated concepts within corpus segments and drill down to term level (1962-1971)

	$Al1k_{S1}$						
Date	1962	1963	1964	1965	1968	1970	1971
Cc_CountThresU_Dim	-	-	-	-	-	-	-
Cv_CountThresU_Dim	Currency	Currency	Currency	Currency	Currency	Currency	Company
	BusinessTerm	BusinessTerm	BusinessTerm	BusinessTerm	BusinessTerm	Company	Currency
	general	general	general	Company	general	general	BusinessTerm
	InsuranceTerm	Geography	InsuranceTerm	general	InsuranceTerm	InsuranceTerm	general
		InsuranceTerm		Geography	Company	BusinessTerm	Geography
				InsuranceTerm	Geography		InsuranceTerm
ELVIA.TermFirstOcc	1995						
ELVIA.TermLastOcc	1998						
Cornhill.TermFirstOcc	1985						
Cornhill.TermLastOcc	1988						
Cc_CountThresU_Company	-	-	-	-	-	-	-
Cv_CountThresU_Company	-	-	-	Union	Globus	Allianz	Allianz
					Lebensversicherungs-AG	Lebensversicherungs-AG	Rechtsschutzversicherungs-AG
						Mercur	
						Rechtsschutzversicherungs-AG	
						Veritas	
	$Al1k_{S2}$						
Date	1962	1963	1964	1965	1968	1970	1971
Cc_CountThresU_Dim	-	-	-	general	-	-	-
Cv_CountThresU_Dim	Currency	Currency	Currency	Currency	Currency	Currency	Currency
	BusinessTerm	BusinessTerm	BusinessTerm	BusinessTerm	BusinessTerm	Company	Company
	general	general	general	general	general	Geography	general
	InsuranceTerm	Geography	InsuranceTerm	InsuranceTerm	InsuranceTerm	InsuranceTerm	Geography
		InsuranceTerm		Company		BusinessTerm	InsuranceTerm
				Geography		general	BusinessTerm
ELVIA.TermFirstOcc	1995						
ELVIA.TermLastOcc	1998						
Cornhill.TermFirstOcc	1986						
Cornhill.TermLastOcc	1998						
Cc_CountThresU_Company	-	-	-	-	-	-	-
Cv_CountThresU_Company	-	-	-	Rueckversicherungs-Gesellschaft	-	Allianz	Allianz
						Allianz-Gruppe	Lebensversicherungs-AG

For the first seven periods from 1962 to 1971 quite similar results were found on aggregated concept level. The occurrence periods of the term “ELVIA” were as expected and exactly the same in both test sets. For “Cornhill” a first occurrence was extracted from “1985” from $Al1k_{S1}$ and “1986” from $Al1k_{S2}$, which matched the expected results to a large extent. The extracted concepts of the other periods are shown in detail in the following tables:

Table 26: Al1k, leading aggregated concepts within corpus segments and drill down to term level (1981-1988)

	$Al1k_{S1}$						
Date	1972	1973	1974	1975	1977	1978	1980
Cc_CountThresU_Dim	-	-	-	-	-	-	-
Cv_CountThresU_Dim	Currency	Currency	Currency	Currency	Currency	Currency	Currency
	Company	Company	Company	Company	Company	Company	Company
	InsuranceTerm	InsuranceTerm	general	InsuranceTerm	BusinessTerm	BusinessTerm	BusinessTerm
	general	general	BusinessTerm	BusinessTerm	general	general	general
	BusinessTerm	BusinessTerm	InsuranceTerm	general	Geography	Geography	Geography
	Geography	Geography		Geography	InsuranceTerm	InsuranceTerm	InsuranceTerm
Cc_CountThresU_Company	-	-	-	-	-	-	-
Cv_CountThresU_Company	Allianz	Allianz	Allianz	Allianz	Allianz	Allianz	Allianz
	Lebensversicherungs-AG	Lebensversicherungs-AG	Lebensversicherungs-AG	Allianz-Sachgruppe			
		Allianz-Gesellschaften					
		Ver-sicherungs-AG					

	$Al1k_{S2}$						
Date	1972	1973	1974	1975	1977	1978	1980
Cc_CountThresU_Dim	-	-	-	-	-	-	-
Cv_CountThresU_Dim	Currency	Company	Currency	Currency	Currency	Currency	Currency
	Company	Currency	Company	Company	Company	Company	Company
	InsuranceTerm	BusinessTerm	BusinessTerm	BusinessTerm	BusinessTerm	general	BusinessTerm
	BusinessTerm	general	general	general	general	InsuranceTerm	general
	general	InsuranceTerm	InsuranceTerm	Geography	Geography	BusinessTerm	Geography
	Geography	Geography	Geography	InsuranceTerm	InsuranceTerm	Geography	InsuranceTerm
Cc_CountThresU_Company	-	-	-	-	-	-	-
Cv_CountThresU_Company	Allianz	Allianz	Allianz	Allianz	Allianz	Allianz	Allianz
	Rechtsschutzversicherungs-AG	Lebensversicherungs-AG	Rechtsschutzversicherungs-AG				
	Assecuranz-Compagnie	Globus					
	Globus	Rechtsschutzversicherungs-AG					

	$Al1k_{S1}$						
Date	1981	1982	1983	1985	1986	1987	1988
Cc_CountThresU_Dim	-	-	-	-	-	-	-
Cv_CountThresU_Dim	Currency	Currency	Company	Currency	Currency	Currency	Currency
	Company	Company	Currency	Company	Company	Company	Company
	InsuranceTerm	BusinessTerm	Geography	BusinessTerm	InsuranceTerm	BusinessTerm	BusinessTerm
	BusinessTerm	general	BusinessTerm	general	Geography	general	general
	general	Geography	general	Geography	BusinessTerm	Geography	Geography
	Geography	InsuranceTerm	InsuranceTerm	InsuranceTerm	general	InsuranceTerm	InsuranceTerm
						Name	
Cc_CountThresU_Company	-	-	-	-	-	-	-
Cv_CountThresU_Company	Allianz	Allianz	Allianz	Allianz	Allianz	Allianz	Allianz
	Hamburg-Mannheimer			RAS	Allianzs		Cornhill
					Cornhill		RAS
					Sicurtae		

	$Al1k_{S2}$						
Date	1981	1982	1983	1985	1986	1987	1988
Cc_CountThresU_Dim	-	-	-	-	-	-	-
Cv_CountThresU_Dim	Currency	Currency	Currency	Company	Currency	Currency	Currency
	Company	Company	Company	Currency	Company	Company	Company
	BusinessTerm	BusinessTerm	BusinessTerm	general	InsuranceTerm	Geography	BusinessTerm
	general	general	general	Geography	Geography	BusinessTerm	general
	InsuranceTerm	InsuranceTerm	Geography	BusinessTerm	BusinessTerm	general	Geography
	Geography		InsuranceTerm	InsuranceTerm	general	InsuranceTerm	InsuranceTerm
Cc_CountThresU_Company	-	-	-	-	-	-	-
Cv_CountThresU_Company	Allianz	Allianz	Allianz	Allianz	Allianz	Allianz	Allianz
	Hamburg-Mannheimer				Allianzs		Cornhill
	Treuegemeinschaftslebensversicherungsgesellschaft				Cornhill		
					Sicurtae		

Table 27: Al1k, leading aggregated concepts within corpus segments and drill down to term level (1989-1995)

	$Al1k_{S1}$						
Date	1989	1990	1991	1992	1993	1994	1995
Cc_CountThresU_Dim	-	-	-	-	-	-	-
Cv_CountThresU_Dim	Currency	Currency	Company	Currency	Currency	Currency	Currency
	Company	Company	Currency	Company	Company	Company	Company
	BusinessTerm	BusinessTerm	general	BusinessTerm	general	InsuranceTerm	general
	general	general	InsuranceTerm	general	Geography	BusinessTerm	BusinessTerm
	InsuranceTerm	Geography	BusinessTerm	Geography	InsuranceTerm	general	Geography
	Geography	InsuranceTerm	Geography	InsuranceTerm	BusinessTerm	Geography	InsuranceTerm
Cc_CountThresU_Company	-	-	-	-	-	-	-
Cv_CountThresU_Company	Allianz	Allianz	Allianz	Allianz	Allianz	Allianz	Allianz
	Eagle	Allianzs		RAS	Allianzs		
	RAS	RAS					
	Star	Rhin					

	$Al1k_{S2}$						
Date	1989	1990	1991	1992	1993	1994	1995
Cc_CountThresU_Dim	-	-	-	-	-	-	-
Cv_CountThresU_Dim	Company	Company	Currency	Currency	Currency	Currency	Currency
	Currency	Currency	Company	Company	Company	Company	Company
	general	BusinessTerm	general	general	general	BusinessTerm	BusinessTerm
	Geography	general	BusinessTerm	InsuranceTerm	BusinessTerm	general	general
	InsuranceTerm	InsuranceTerm	InsuranceTerm	BusinessTerm	Geography	Geography	Geography
	BusinessTerm	Geography	Geography	Geography	InsuranceTerm	InsuranceTerm	InsuranceTerm
						Name	
Cc_CountThresU_Company	-	-	-	-	-	-	-
Cv_CountThresU_Company	Allianz	Allianz	Allianz	Allianz	Allianz	Allianz	Allianz
	Star	Allianzs	RAS		Allianzs		Adriatico
							ELVIA

Table 28: Al1k, leading aggregated concepts within corpus segments and drill down to term level (1996-2000)

	$Al1k_{S1}$			
Date	1996	1998	1999	2000
Cc_CountThresU_Dim	-	-	-	-
Cv_CountThresU_Dim	Company	Currency	Currency	Company
	Currency	Company	Company	Currency
	BusinessTerm	BusinessTerm	BusinessTerm	general
	general	general	general	Geography
	Geography	Geography	Geography	BusinessTerm
	InsuranceTerm	InsuranceTerm	InsuranceTerm	InsuranceTerm
				Name
Cc_CountThresU_Company	-	-	-	-
Cv_CountThresU_Company	Allianz	Allianz	Allianz	Allianz
		Adriatico	AGF	AGF
		AGF	Hermes	
		Assurances	RAS	
		Datastream		
		ELVIA		
		Federales		

	$Al1k_{S2}$			
Date	1996	1998	1999	2000
Cc_CountThresU_Dim	-	-	-	-
Cv_CountThresU_Dim	Company	Currency	Currency	Company
	Currency	Company	Company	Currency
	Geography	BusinessTerm	BusinessTerm	Geography
	general	general	general	general
	BusinessTerm	InsuranceTerm	Geography	Name
	InsuranceTerm	Geography	InsuranceTerm	BusinessTerm
			Name	InsuranceTerm
Cc_CountThresU_Company	-	-	-	-
Cv_CountThresU_Company	Allianz	Allianz	Allianz	Allianz
	AZT	AGF	AGF	
		ELVIA		
		RAS		

In contrast to quite similar results on aggregated concept level, the results of term level showed very few similarities compared to the test sets. For corpus

segment C_C no concept or term was extracted during all periods. This was the case probably due to corpus size limitations.

4.3.1.6 Semantic analysis of type n corpora summary

- *The extracted knowledge was more precise the more tokens the test set contained.*
- *Concepts were successfully extracted which were pre-selected as exemplary benchmarks from external domain-expert background knowledge.*
- *CW_{5k} and CW_{1k} reflected comparable knowledge when analyzed with the segmenting approach with TRQ threshold.*

4.3.2 Evaluation of Corpus type b (low pre-processing intensity)

In this chapter test sets are analysed which do not match the optimal case of data with $C_G = 0$. Here corpora are analysed with $C_G \neq 0$.

4.3.2.1 Statistical analysis of type b corpora

According to the procedure in Chapter 4.3.1.1 in the following sub-chapters the statistical qualities of all test sets are introduced.

4.3.2.1.1 Descriptive statistics of CW corpus test set CW_{5kb}

An overview of statistical qualities of the corpus C, the corpus segments C_C and C_V based on TRQ measures is documented in Table 57 (see Appendix).

The absolute value of TRQ is high within C_C and low within C_V. This is also the case for mean and standard deviation. The skewness is positive for all segments. The value of kurtosis is always positive. The distribution of values does not vary between the different vertical corpus segments. This is a different result compared to that derived from CW_{5k} test set (see Table 48). Espe-

cially interesting is that the negative skewness for C_C and the negative kurtosis for C_C and C_V were not found.

Table 29: Correlations between corpus segments based on TRQ measure

Correlations

		Cw5kbCRepQuo	Cw5kbCcRepQuo	Cw5kbCvRepQuo
Cw5kbCRepQuo	Pearson Correlation	1	,987**	,943**
	Sig. (2-tailed)		,000	,000
	Sum of Squares and Cross-products	83067,313	998878,481	8900,963
	Covariance	2966,690	35674,231	317,892
	N	29	29	29
Cw5kbCcRepQuo	Pearson Correlation	,987**	1	,883**
	Sig. (2-tailed)	,000		,000
	Sum of Squares and Cross-products	998878,481	12336286,729	101517,335
	Covariance	35674,231	440581,669	3625,619
	N	29	29	29
Cw5kbCvRepQuo	Pearson Correlation	,943**	,883**	1
	Sig. (2-tailed)	,000	,000	
	Sum of Squares and Cross-products	8900,963	101517,335	1071,542
	Covariance	317,892	3625,619	38,269
	N	29	29	29

**. Correlation is significant at the 0.01 level (2-tailed).

In Table 29 significant, strong, positive correlations can be seen (based on TRQ measure) between C and C_V as well as C_C corpus segments. This is a significant difference to the results found for corpus test sets with $C_G = 0$. The segmentation of C into a constant and a volatile segment did not lead to statistically significant different corpus segments.

4.3.2.1.2 Descriptive statistics of CW corpus test set CW_{5kbu}

An overview of statistical qualities of the corpus C, the corpus segments C_C and C_V based on TRQ measures is documented in Table 58 (see Appendix).

The absolute value of TRQ is high within C_C and low within C_V. This is also the case for mean and standard deviation. The skewness is positive for all segments. The value of kurtosis is always positive. The distribution of values does not vary among the different vertical corpus segments.

Table 30: Correlations between corpus segments based on TRQ measure

Correlations

		Cw5kbu CRepQuo	Cw5kbuCc RepQuo	Cw5kbuCv RepQuo
Cw5kbuCRepQuo	Pearson Correlation	1	,985**	,940**
	Sig. (2-tailed)		,000	,000
	Sum of Squares and Cross-products	77004,510	822250,612	8125,427
	Covariance	2750,161	29366,093	290,194
	N	29	29	29
Cw5kbuCcRepQuo	Pearson Correlation	,985**	1	,877**
	Sig. (2-tailed)	,000		,000
	Sum of Squares and Cross-products	822250,612	9044478,638	82177,355
	Covariance	29366,093	323017,094	2934,906
	N	29	29	29
Cw5kbuCvRepQuo	Pearson Correlation	,940**	,877**	1
	Sig. (2-tailed)	,000	,000	
	Sum of Squares and Cross-products	8125,427	82177,355	969,995
	Covariance	290,194	2934,906	34,643
	N	29	29	29

**. Correlation is significant at the 0.01 level (2-tailed).

In Table 30 significant strong positive correlations can be found (based on TRQ measure) between C and C_V as well as C_C corpus segments. This is significantly different from the results found for corpus test sets with $C_G = 0$. The segmentation of C into a constant and a volatile segment did not lead to statistically significant different corpus segments.

4.3.2.1.3 Descriptive statistics of CW corpus test set CW_{5kbun} and CW_{5kbun2}

An overview of statistical qualities of the corpus C, the corpus segments C_C and C_V based on TRQ measures is documented in Table 59 and Table 60 (see Appendix).

The absolute value of TRQ is high within C_C and low within C_V. This is also the case for mean and standard deviation. The skewness is positive for C and C_V, but negative for C_C. That is comparable to the test sets with $C_G = 0$ (see chapter Statistical analysis) and seems to be dependent on the even size of yearly corpus segments in test set CW_{5kbun}. The value of kurtosis is always positive. The distribution of values does not vary between the different vertical corpus segments.

Table 31: Correlations between corpus segments based on TRQ measure

Correlations

		Cw5kbun CRepQuo	Cw5kbunCc RepQuo	Cw5kbunCv RepQuo
Cw5kbunCRepQuo	Pearson Correlation	1	,194	,943**
	Sig. (2-tailed)		,314	,000
	Sum of Squares and Cross-products	4406,668	334,449	611,338
	Covariance	157,381	11,945	21,833
	N	29	29	29
Cw5kbunCcRepQuo	Pearson Correlation	,194	1	-,142
	Sig. (2-tailed)	,314		,464
	Sum of Squares and Cross-products	334,449	676,055	-35,961
	Covariance	11,945	24,145	-1,284
	N	29	29	29
Cw5kbunCvRepQuo	Pearson Correlation	,943**	-,142	1
	Sig. (2-tailed)	,000	,464	
	Sum of Squares and Cross-products	611,338	-35,961	95,300
	Covariance	21,833	-1,284	3,404
	N	29	29	29

**. Correlation is significant at the 0.01 level (2-tailed).

Table 32: Correlations between corpus segments based on TRQ measure

Correlations

		Cw5kbun2 CRepQuo	Cw5kbun2 CcRepQuo	Cw5kbun2 CvRepQuo
Cw5kbun2CRepQuo	Pearson Correlation	1	,068	,935**
	Sig. (2-tailed)		,727	,000
	Sum of Squares and Cross-products	2928,051	84,367	424,452
	Covariance	104,573	3,013	15,159
	N	29	29	29
Cw5kbun2CcRepQuo	Pearson Correlation	,068	1	-,291
	Sig. (2-tailed)	,727		,126
	Sum of Squares and Cross-products	84,367	529,437	-56,095
	Covariance	3,013	18,908	-2,003
	N	29	29	29
Cw5kbun2CvRepQuo	Pearson Correlation	,935**	-,291	1
	Sig. (2-tailed)	,000	,126	
	Sum of Squares and Cross-products	424,452	-56,095	70,425
	Covariance	15,159	-2,003	2,515
	N	29	29	29

**. Correlation is significant at the 0.01 level (2-tailed).

In Table 31 and Table 32 significant strong positive correlations can be found (based on TRQ measure) among C and C_V corpus segments. In contrast to previous test sets with $C_G \neq 0$ C_C and C_V are not positively correlated, but

the found negative correlation is weak and not as significant as for corpus test sets with $C_G = 0$. The segmentation of C into a constant and a volatile segment again did not lead to statistically significant different corpus segments.

4.3.2.1.4 Descriptive statistics of CW corpus test set CW_{1kb}

An overview of statistical qualities of the corpus C, the corpus segments C_C and C_V based on TRQ measures is documented in Table 61 (see Appendix).

The absolute value of TRQ is high within C_C and low within C_V. This is also the case for mean and standard deviation. The skewness is positive for all segments. The value of kurtosis is always positive. The distribution of values does not vary among the different vertical corpus segments.

Table 33: Correlations between corpus segments based on TRQ measure

Correlations

		Cw1kbCRepQuo	Cw1kbCcRepQuo	Cw1kbCvRepQuo
Cw1kbCRepQuo	Pearson Correlation	1	,991**	,958**
	Sig. (2-tailed)		,000	,000
	Sum of Squares and Cross-products	35407,066	402477,263	4042,814
	Covariance	1264,538	14374,188	144,386
	N	29	29	29
Cw1kbCcRepQuo	Pearson Correlation	,991**	1	,916**
	Sig. (2-tailed)	,000		,000
	Sum of Squares and Cross-products	402477,263	4655600,670	44317,865
	Covariance	14374,188	166271,452	1582,781
	N	29	29	29
Cw1kbCvRepQuo	Pearson Correlation	,958**	,916**	1
	Sig. (2-tailed)	,000	,000	
	Sum of Squares and Cross-products	4042,814	44317,865	502,499
	Covariance	144,386	1582,781	17,946
	N	29	29	29

**. Correlation is significant at the 0.01 level (2-tailed).

In Table 33 significant strong positive correlations can be found (based on TRQ measure) between C and C_V as well as C_C corpus segments. This is significantly different to the results found for corpus test sets with $C_G = 0$. The segmentation of C into a constant and volatile segment did not lead to statistically significant different corpus segments.

4.3.2.1.5 Descriptive statistics of CW corpus test set CW_{1kbu}

An overview of statistical qualities of the corpus C, the corpus segments C_C and C_V based on TRQ measures is documented in Table 62 (see Appendix).

The absolute value of TRQ is high within C_C and low within C_V. This is also the case for mean and standard deviation. The skewness is positive for all segments. The value of kurtosis is always positive. The distribution of values does not vary among the different vertical corpus segments.

Table 34: Correlations between corpus segments based on TRQ measure

Correlations

		Cw1kbu CRepQuo	Cw1kbuCc RepQuo	Cw1kbuCv RepQuo
Cw1kbuCRepQuo	Pearson Correlation	1	,991**	,959**
	Sig. (2-tailed)		,000	,000
	Sum of Squares and Cross-products	33671,622	351288,276	3768,485
	Covariance	1202,558	12546,010	134,589
	N	29	29	29
Cw1kbuCcRepQuo	Pearson Correlation	,991**	1	,916**
	Sig. (2-tailed)	,000		,000
	Sum of Squares and Cross-products	351288,276	3733887,172	37931,501
	Covariance	12546,010	133353,113	1354,696
	N	29	29	29
Cw1kbuCvRepQuo	Pearson Correlation	,959**	,916**	1
	Sig. (2-tailed)	,000	,000	
	Sum of Squares and Cross-products	3768,485	37931,501	458,888
	Covariance	134,589	1354,696	16,389
	N	29	29	29

**. Correlation is significant at the 0.01 level (2-tailed).

In Table 34 significant strong positive correlations can be found (based on TRQ measure) between C and C_V as well as C_C corpus segments. This is significantly different from the results found for corpus test sets with $C_G = 0$. The segmentation of C into a constant and volatile segment did not lead to statistically significant different corpus segments.

4.3.2.2 Statistical analysis of type b corpora summary

Compared to type “n” corpora (see Table 17) the type “b” corpora led to different results. Especially the significant negative correlation between C_C and C_V was not found in the type “b” corpus test sets.

Table 35: Statistical analysis summary corpus type b

Test Set	Correlation of corpus segment pair C-C_C	Correlation of corpus segment pair C-C_V	Correlation of corpus segment pair C_C-C_V
CW_{5kb}	Positive, strong, significant	Positive, strong, significant	Positive, strong, significant
CW_{5kbu}	Positive, strong, significant	Positive, strong, significant	Positive, strong, significant
CW_{5kbun}	Positive, weak	Positive, strong, significant	Negative, weak
CW_{5kbun2}	Positive, weak	Positive, strong, significant	Negative, weak
CW_{1kb}	Positive, strong, significant	Positive, strong, significant	Positive, strong, significant
CW_{1kbu}	Positive, strong, significant	Positive, strong, significant	Positive, strong, significant

In the test sets CW_{5kbun} and CW_{5kbun2} the corpus length dependency of TRQ was considered by normalizing the yearly time segments of corpora to a common amount of terms. But this did not overlay the absence of the typically significant correlation between C_C and C_V in type “n” corpora.

4.3.2.3 Distribution analysis of applied taxonomies on type b corpora

Results in this chapter will be evaluated by a comparison with results derived from corpus type “n” (documented in Chapter 4.3.1.3). Table 52 showed very typical shapes for the graphs from number of yearly matched terms by “Dim” taxonomies for type “n” corpora for segments C_V and C_C. Major differences between these two segments are not visible in Table 63 (see Appendix), neither the S-shape for C_V corpus segments, nor the quite even distribution for C_C concepts.

Due to the less prevalent number of “Dim_Mertens” concepts, compared to type “n” corpora; the absolute number of assigned concepts was lower. Differences between C and CV corpus segments were not found. For the more limited type “bn” corpora CW_{5kbun} and CW_{5kbun2} again no concepts matched with “Dim_Mertens” taxonomy. A discussion of found results per test set follows in the next chapters.

4.3.2.3.1 Distribution Analysis of CW corpus test set CW_{5kb}

In this chapter the statistical qualities of the resulting corpus data sets are analysed, when different taxonomies are applied to CW_{5kb}.

In contrast to

Table 54 in Table 64 (see Appendix) the number of assigned terms to the “Dim” taxonomy is very volatile with a factor of approx. 2.5. The typical graph with an “S”-like shape was again not found. Variance and standard deviation also appear with higher values. For the other corpus segments a very volatile assignment of terms was also found.

The “Dim_Mertens” taxonomy volatile assignment of terms was also present for the CW_{5k} test set (see

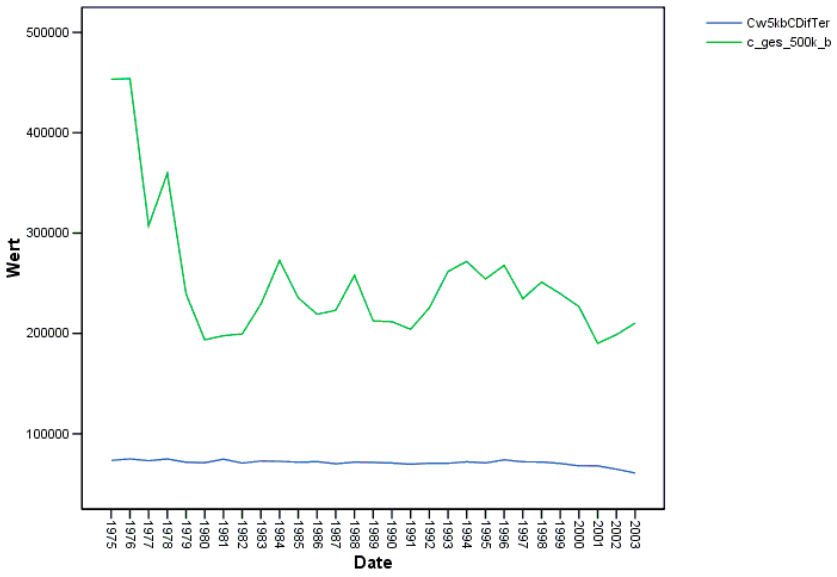
Table 54). For CW_{5kb} the volatility rose once more.

The assignment of terms to “CC_Dim” delivers values that indicate a higher volatility of test set time series for the descriptive statistic measures. The filtering effect of applied taxonomies was lower than that found for the CW_{5k} test set.

Compared to CW_{5k} results (see

Table 18) significant, negative correlations among the number of assigned terms to the taxonomies applied to segments C_V and C were not present for CW_{5kb} (see Table 36). Conversely for CW_{5kb} positive correlations were found within this analysis, with a significant result for C.

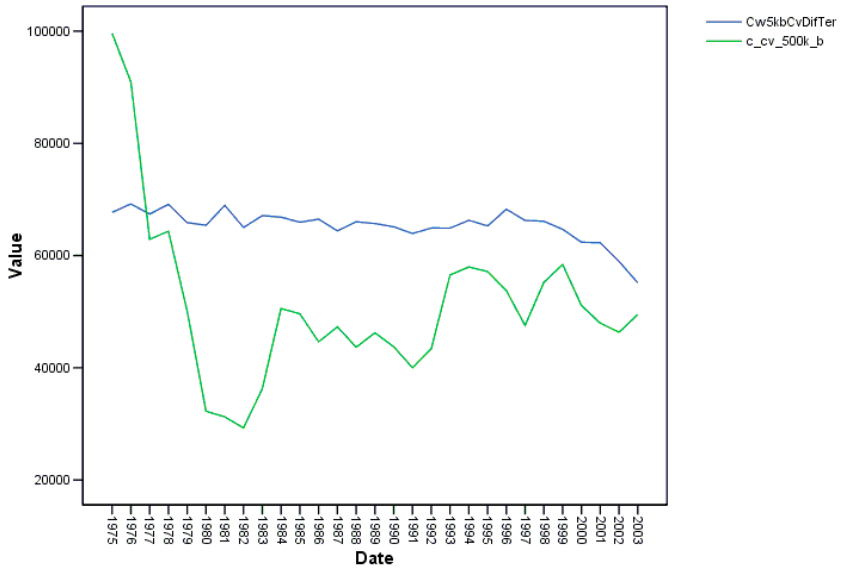
Table 36: Analysis of statistical dependence between the TRQ graph and the graph of assigned terms by the Dim taxonomy for CW_{5kb}

Segment	Results of correlation analysis
C	

Korrelationen

		Cw5kbCDif Ter	c_ges_ 500k_b
Cw5kbCDifTer	Korrelation nach Pearson	1	,499**
	Signifikanz (2-seitig)		,006
	Quadratsummen und Kreuzprodukte	237559687,0	2,7E+09
	Kovarianz	8484274,534	9,7E+07
	N	29	29
c_ges_500k_b	Korrelation nach Pearson	,499**	1
	Signifikanz (2-seitig)	,006	
	Quadratsummen und Kreuzprodukte	2714241087	1,2E+11
	Kovarianz	96937181,677	4,5E+09
	N	29	29

**. Die Korrelation ist auf dem Niveau von 0,01 (2-seitig) signifikant.

Segment	Results of correlation analysis
C_V	

Korrelationen

		Cw5kbCvDifTer	c_cv_500k_b
Cw5kbCvDifTer	Korrelation nach Pearson	1	,262
	Signifikanz (2-seitig)		,169
	Quadratsummen und Kreuzprodukte	237559687,0	321213462,3
	Kovarianz	8484274,534	11471909,37
	N	29	29
c_cv_500k_b	Korrelation nach Pearson	,262	1
	Signifikanz (2-seitig)	,169	
	Quadratsummen und Kreuzprodukte	321213462,3	6313725856
	Kovarianz	11471909,368	225490209,2
	N	29	29

4.3.2.3.2 Distribution Analysis of CW corpus test set CW_{5kbu}

In this chapter the statistical qualities of the resulting corpus data sets are analysed when different taxonomies are applied to CW_{5kbu}.

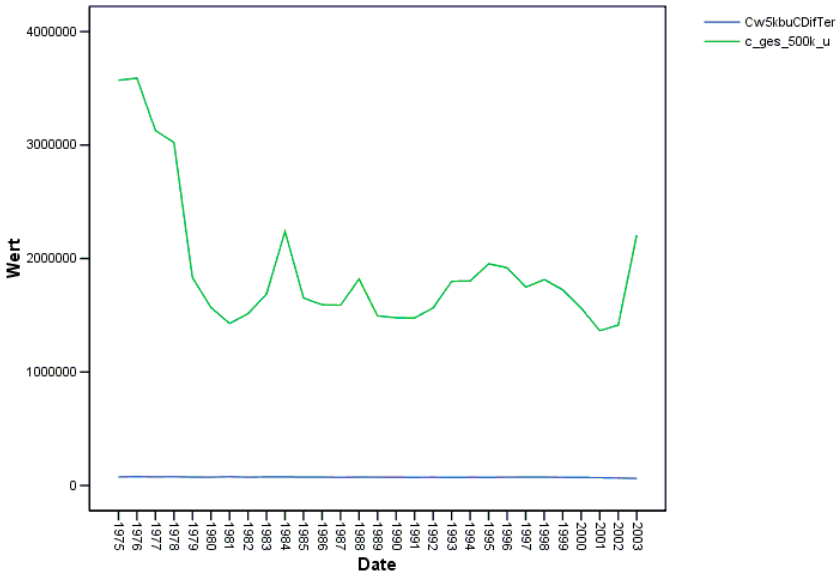
Comparable with test set CW_{5k} and contrary to

Table 54, in Table 65 (see Appendix) the number of assigned terms to the “Dim” taxonomy is very volatile with a factor of approx. 2.5. The typical graph with an “S”-like shape was again not found. Variance and standard deviation

also appear with higher values. Very volatile assignments of terms were also found for the other corpus segments.

The “Dim_Mertens” taxonomy volatile assignments of terms and the filtering effect of terms when taxonomy “CC_Dim” is applied were again present with similar results to the CW_{5kb} test set.

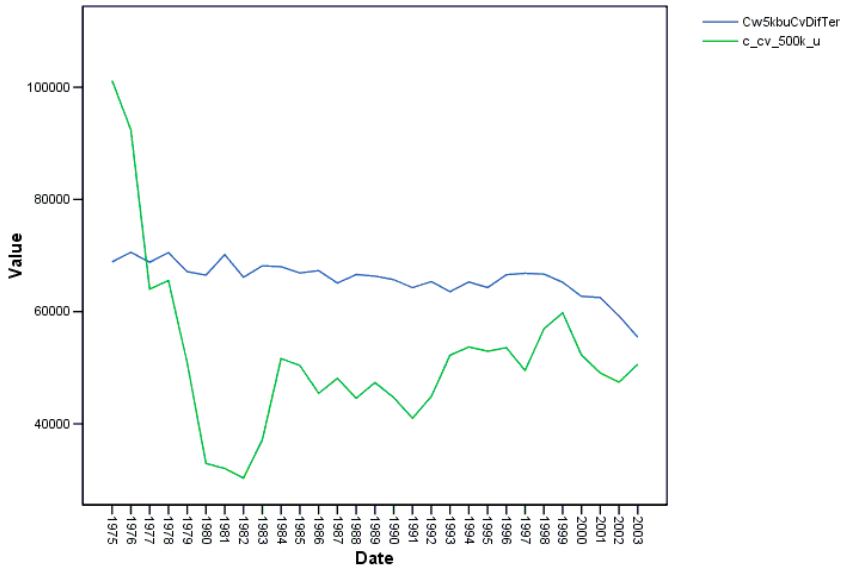
Table 37: Analysis of statistical dependence between the TRQ graph and the graph of assigned terms by the Dim taxonomy for CW_{5kbu}

Segment	Results of correlation analysis
C	

Korrelationen

		Cw5kbuCDifTer	c_ges_500k_u
Cw5kbuCDifTer	Korrelation nach Pearson	1	,412*
	Signifikanz (2-seitig)		,026
	Quadratsummen und Kreuzprodukte	284021955,2	2,3E+10
	Kovarianz	10143641,259	8,1E+08
	N	29	29
c_ges_500k_u	Korrelation nach Pearson	,412*	1
	Signifikanz (2-seitig)	,026	
	Quadratsummen und Kreuzprodukte	22778632413	1,1E+13
	Kovarianz	813522586,2	3,8E+11
	N	29	29

*. Die Korrelation ist auf dem Niveau von 0,05 (2-seitig) signifikant.

Segment	Results of correlation analysis
C_V	

Korrelationen

		Cw5kbuCv DifTer	c_cv_500k_u
Cw5kbuCvDifTer	Korrelation nach Pearson	1	,262
	Signifikanz (2-seitig)		,169
	Quadratsummen und Kreuzprodukte	284021955,2	351293383,8
	Kovarianz	10143641,259	12546192,28
	N	29	29
c_cv_500k_u	Korrelation nach Pearson	,262	1
	Signifikanz (2-seitig)	,169	
	Quadratsummen und Kreuzprodukte	351293383,8	6321880405
	Kovarianz	12546192,277	225781443,0
	N	29	29

Compared to CW_{5k} results (see

Table 18) the significant negative correlations between the number of assigned terms to the taxonomies applied to segments C_V and C were not present for CW_{5kbu} (see Table 37). Quite the contrary, for CW_{5kbu} positive correlations were found within this analysis with a significant result for C.

These results are quite similar to those found for CW_{5kb} with lower intensity of pre-processing.

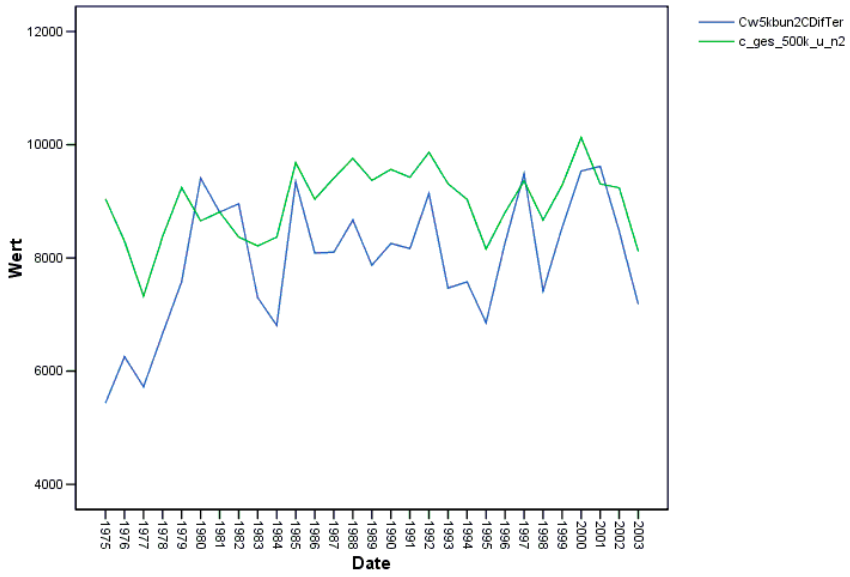
4.3.2.3.3 Distribution Analysis of CW corpus test sets CW_{5kbun} and CW_{5kbun2}

In this chapter the statistical qualities of the resulting corpus data sets are analysed, when different taxonomies are applied to CW_{5kbun} and CW_{5kbun2}.

Table 38: Analysis of statistical dependence between the TRQ graphs and the graphs of assigned terms by the Dim taxonomy for CW_{5kbun} and CW_{5kbun2}

Corpus	Segment	Results of correlation analysis
CW_{5kbun}	C	

Corpus	Segment	Results of correlation analysis
		Korrelationen **Cw5kbunCDifTer / c_ges_500k_u_n** Cw5kbunCDifTer: Korrelation nach Pearson: 1 / ,631** Signifikanz (2-seitig): / ,000 Quadratsummen und Kreuzprodukte: 42757796,690 / 13956466 Kovarianz: 1527064,167 / 498445,22 N: 29 / 29 c_ges_500k_u_n: Korrelation nach Pearson: ,631** / 1 Signifikanz (2-seitig): ,000 / Quadratsummen und Kreuzprodukte: 13956466,034 / 11429111 Kovarianz: 498445,216 / 408182,52 N: 29 / 29 **. Die Korrelation ist auf dem Niveau von 0,01 (2-seitig) signifikant.
	C_V	
		Korrelationen **Cw5kbunCvDifTer / c_cv_500k_u_n** Cw5kbunCvDifTer: Korrelation nach Pearson: 1 / ,656** Signifikanz (2-seitig): / ,000 Quadratsummen und Kreuzprodukte: 42757796,690 / 12036130 Kovarianz: 1527064,167 / 429861,79 N: 29 / 29 c_cv_500k_u_n: Korrelation nach Pearson: ,656** / 1 Signifikanz (2-seitig): ,000 / Quadratsummen und Kreuzprodukte: 12036130,034 / 7867324,6 Kovarianz: 429861,787 / 280975,88 N: 29 / 29 **. Die Korrelation ist auf dem Niveau von 0,01 (2-seitig) signifikant.

Corpus	Segment	Results of correlation analysis
CW_{5kbun2}	C	
		Korrelationen **. Die Korrelation ist auf dem Niveau von 0,01 (2-seitig) signifikant.

		Cw5kbun2 CDifTer	c_ges_ 500k_u_n2
Cw5kbun2CDifTer	Korrelation nach Pearson	1	,642**
	Signifikanz (2-seitig)		,000
	Quadratsummen und Kreuzprodukte	36425708,966	13003142,9
	Kovarianz	1300918,177	464397,962
	N	29	29
c_ges_500k_u_n2	Korrelation nach Pearson	,642**	1
	Signifikanz (2-seitig)	,000	
	Quadratsummen und Kreuzprodukte	13003142,931	11276485,9
	Kovarianz	464397,962	402731,638
	N	29	29

	C_V	

Corpus	Segment	Results of correlation analysis

Korrelationen

		Cw5kbun2 CvDifTer	c_cv_500k_ u_n2
Cw5kbun2CvDifTer	Korrelation nach Pearson	1	,634**
	Signifikanz (2-seitig)		,000
	Quadratsummen und Kreuzprodukte	36425708,966	9479682,310
	Kovarianz	1300918,177	338560,083
	N	29	29
c_cv_500k_u_n2	Korrelation nach Pearson	,634**	1
	Signifikanz (2-seitig)	,000	
	Quadratsummen und Kreuzprodukte	9479682,310	6139174,207
	Kovarianz	338560,083	219256,222
	N	29	29

**. Die Korrelation ist auf dem Niveau von 0,01 (2-seitig) signifikant.

Compared to CW_{5k} results (see

Table 18) the significant negative correlations between the number of assigned terms to the taxonomies applied with segments C_V and C were also not present for CW_{5kbun} and CW_{5kbun2} (see Table 38 in Appendix). To a greater extent than for CW_{5kb}, for CW_{5kbun} and CW_{5kbun2} significant, positive correlations were found within this analysis.

These results are quite similar to those found for CW_{5kbu}. The applied quantitative normalizing of yearly corpus segments did not overlay the present test set internal statistical qualities.

4.3.2.4 Distribution analysis on type b corpora summary

On the contrary from the results from type “n” corpora (see Chapter 4.3.1.4), no negative correlation was found among the number of different terms and the number of terms assigned to a given taxonomy for each yearly segment (see Table 39). This correlation was either weak or significant, depending on the intensity of pre-processing.

Table 39: Correlation between types and assigned terms per year by taxonomy within corpus type b

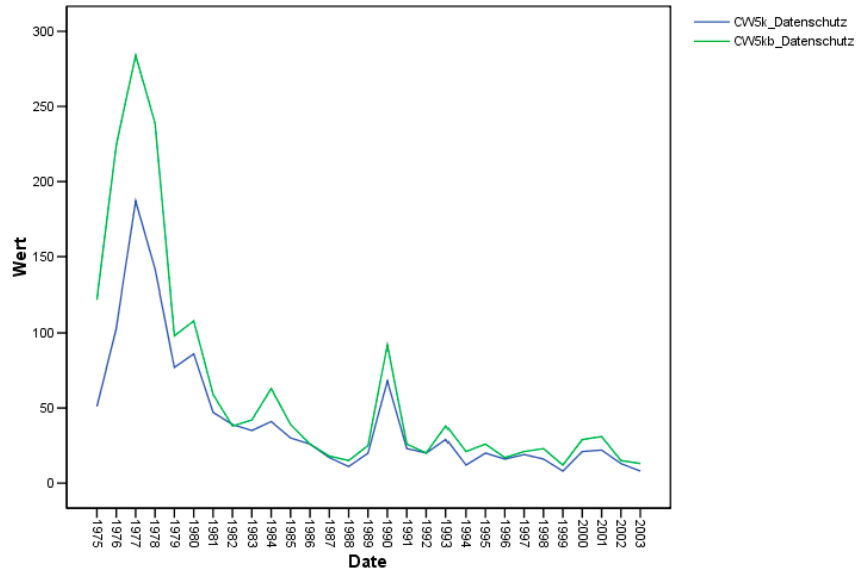
Test Set	Taxonomy	Corpus segment C	Corpus segment C_V
CW_{5kb}	Dim	Positive, middle, significant	Positive, middle
CW_{5kbu}	Dim	Positive, middle, significant	Positive, middle
CW_{5kbun}	Dim	Positive, middle, significant	Positive, middle, significant
CW_{5kbun2}	Dim	Positive, middle, significant	Positive, middle, significant

4.3.2.5 Semantic analysis of type b corpora

In this chapter the results found for corpus type “n” test sets (see Chapter 4.3.1.5) are compared to knowledge extracted from corpus type “b” and “bn” test sets. A summary is given in Chapter 4.3.2.6.

4.3.2.5.1 Semantic analysis of CW corpus test set CW_{5kb}

In the following diagrams the progress paths of certain concepts from the “Dim_Mertens” taxonomy extracted from the CW_{5kb} corpus test set are analysed regarding their correlation to that extracted from CW_{5k}. The correlations among all concepts compared were strong and significant at 99% (both sided). For these pre-selected concepts the lower intensity of pre-processing did not lead to significantly different extracted progress paths.

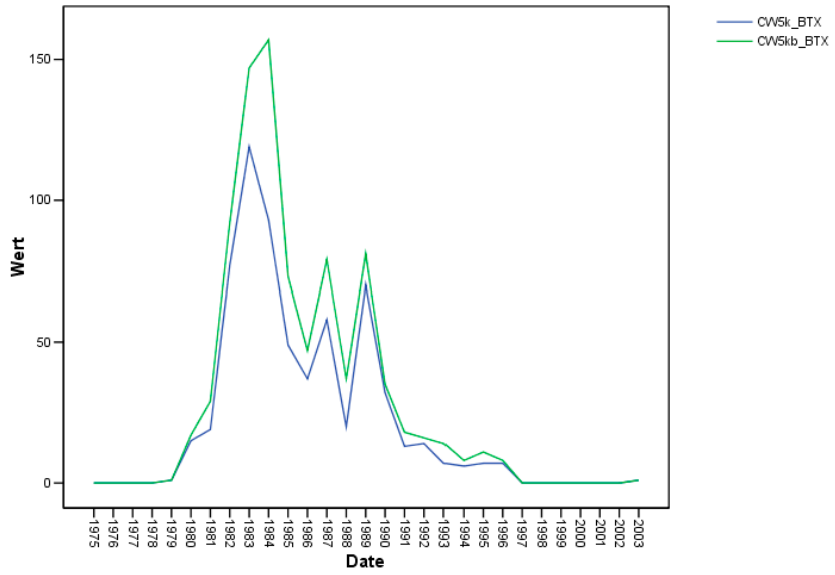


Korrelationen

		CW5k_ Datenschutz	CW5kb_ Datenschutz
CW5k_Datenschutz	Korrelation nach Pearson	1	,970**
	Signifikanz (2-seitig)		,000
	Quadratsummen und Kreuzprodukte	49486,828	81900,690
	Kovarianz	1767,387	2925,025
	N	29	29
CW5kb_Datenschutz	Korrelation nach Pearson	,970**	1
	Signifikanz (2-seitig)	,000	
	Quadratsummen und Kreuzprodukte	81900,690	144087,241
	Kovarianz	2925,025	5145,973
	N	29	29

**. Die Korrelation ist auf dem Niveau von 0,01 (2-seitig) signifikant.

Fig. 51: Correlation between progresses of the concept "Datenschutz" extracted from CW_{5k} and from CW_{5kb} corpus

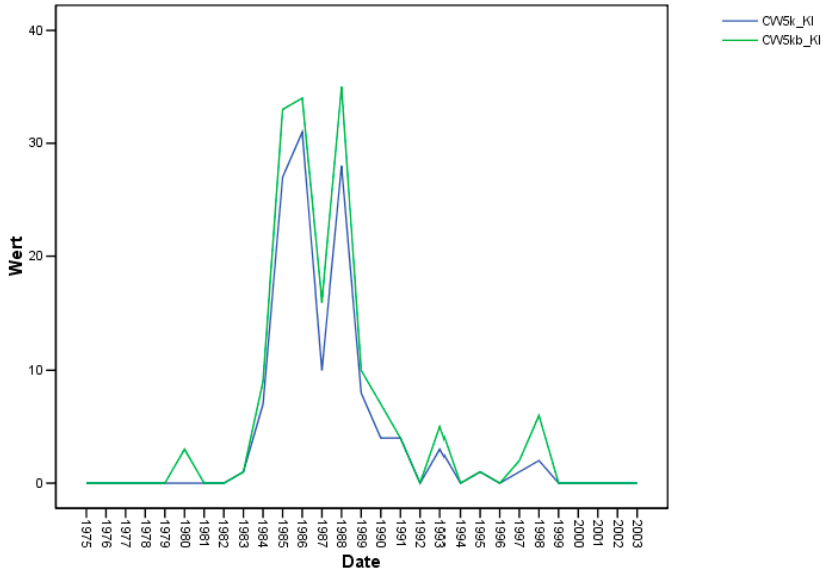


Korrelationen

		CW5k_BTX	CW5kb_BTX
CW5k_BTX	Korrelation nach Pearson	1	,983**
	Signifikanz (2-seitig)		,000
	Quadratsummen und Kreuzprodukte	28987,310	38778,759
	Kovarianz	1035,261	1384,956
	N	29	29
CW5kb_BTX	Korrelation nach Pearson	,983**	1
	Signifikanz (2-seitig)	,000	
	Quadratsummen und Kreuzprodukte	38778,759	53652,966
	Kovarianz	1384,956	1916,177
	N	29	29

**. Die Korrelation ist auf dem Niveau von 0,01 (2-seitig) signifikant.

Fig. 52: Correlation between progresses of the concept "BTX" extracted from CW_{5k} and from CW_{5kb} corpus

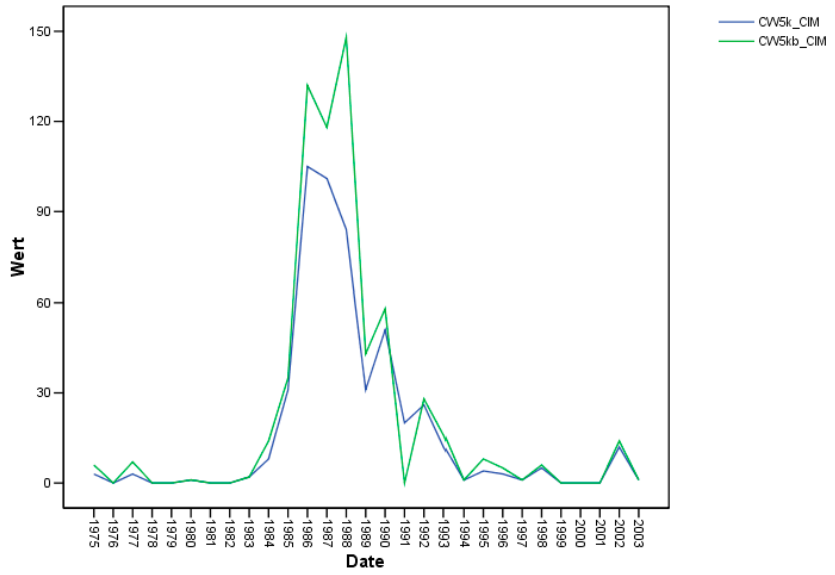


Korrelationen

		CW5k_KI	CW5kb_KI
CW5k_KI	Korrelation nach Pearson	1	,992**
	Signifikanz (2-seitig)		,000
	Quadratsummen und Kreuzprodukte	2178,828	2576,034
	Kovarianz	77,815	92,001
	N	29	29
CW5kb_KI	Korrelation nach Pearson	,992**	1
	Signifikanz (2-seitig)	,000	
	Quadratsummen und Kreuzprodukte	2576,034	3097,793
	Kovarianz	92,001	110,635
	N	29	29

**. Die Korrelation ist auf dem Niveau von 0,01 (2-seitig) signifikant.

Fig. 53: Correlation between progresses of the concept "KI" extracted from CW_{5k} and from CW_{5kb} corpus

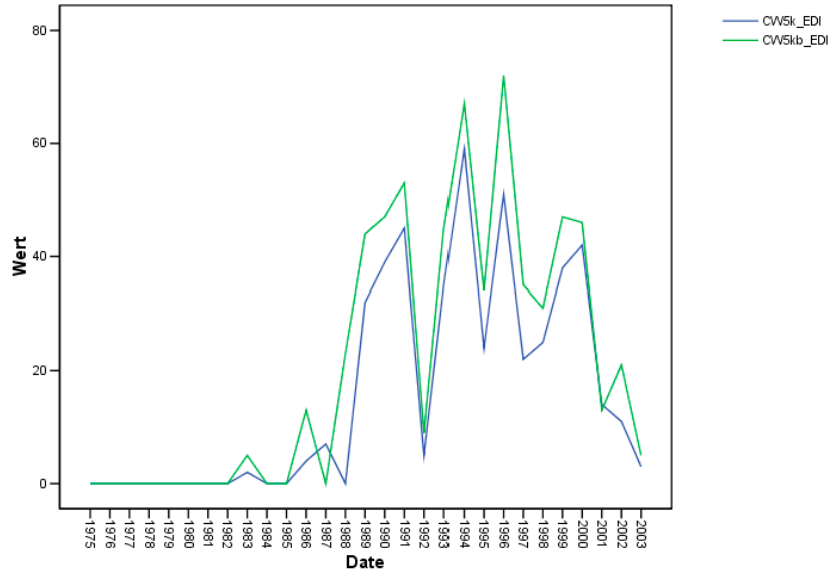


Korrelationen

		CW5k_CIM	CW5kb_CIM
CW5k_CIM	Korrelation nach Pearson	1	,973**
	Signifikanz (2-seitig)		,000
	Quadratsummen und Kreuzprodukte	25515,034	33695,517
	Kovarianz	911,251	1203,411
	N	29	29
CW5kb_CIM	Korrelation nach Pearson	,973**	1
	Signifikanz (2-seitig)	,000	
	Quadratsummen und Kreuzprodukte	33695,517	47038,759
	Kovarianz	1203,411	1679,956
	N	29	29

**. Die Korrelation ist auf dem Niveau von 0,01 (2-seitig) signifikant.

Fig. 54: Correlation between progresses of the concept "CIM" extracted from CW_{5k} and from CW_{5kb} corpus

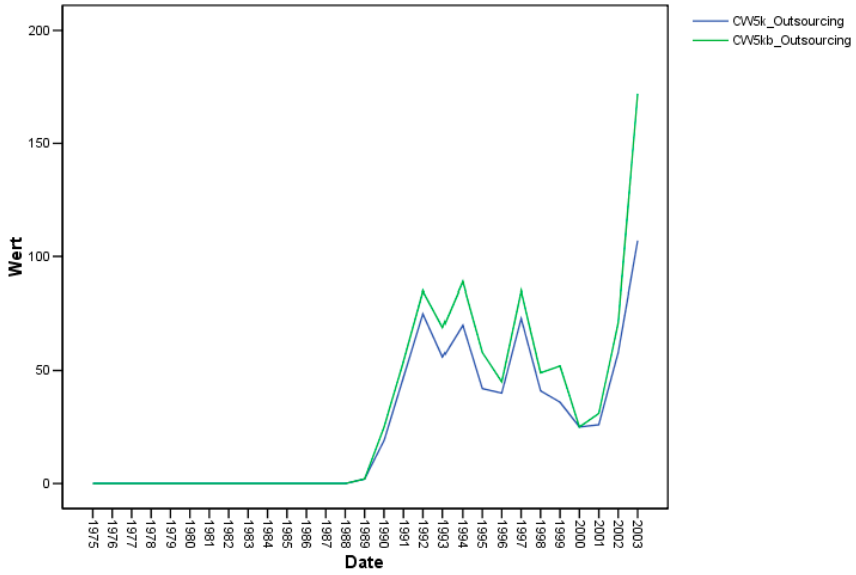


Korrelationen

		CW5k_EDI	CW5kb_EDI
CW5k_EDI	Korrelation nach Pearson	1	,968**
	Signifikanz (2-seitig)		,000
	Quadratsummen und Kreuzprodukte	9956,759	11806,207
	Kovarianz	355,599	421,650
	N	29	29
CW5kb_EDI	Korrelation nach Pearson	,968**	1
	Signifikanz (2-seitig)	,000	
	Quadratsummen und Kreuzprodukte	11806,207	14926,966
	Kovarianz	421,650	533,106
	N	29	29

**. Die Korrelation ist auf dem Niveau von 0,01 (2-seitig) signifikant.

Fig. 55: Correlation between progresses of the concept "EDI" extracted from CW_{5k} and from CW_{5kb} corpus



Korrelationen

		CW5k_ Outsourcing	CW5kb_ Outsourcing
CW5k_Outsourcing	Korrelation nach Pearson	1	,984**
	Signifikanz (2-seitig)		,000
	Quadratsummen und Kreuzprodukte	26291,793	35212,586
	Kovarianz	938,993	1257,592
	N	29	29
CW5kb_Outsourcing	Korrelation nach Pearson	,984**	1
	Signifikanz (2-seitig)	,000	
	Quadratsummen und Kreuzprodukte	35212,586	48701,172
	Kovarianz	1257,592	1739,328
	N	29	29

**. Die Korrelation ist auf dem Niveau von 0,01 (2-seitig) signifikant.

Fig. 56: Correlation between progresses of the concept "Outsourcing" extracted from CW_{5k} and from CW_{5kb} corpus

Korrelationen

		CW5k_ Dezentralisierung	CW5kb_ Dezentralisierung
CW5k_Dezentralisierung	Korrelation nach Pearson	1	,934**
	Signifikanz (2-seitig)		,000
	Quadratsummen und Kreuzprodukte	2301,310	3253,897
	Kovarianz	82,190	116,211
	N	29	29
CW5kb_ Dezentralisierung	Korrelation nach Pearson	,934**	1
	Signifikanz (2-seitig)	,000	
	Quadratsummen und Kreuzprodukte	3253,897	5277,034
	Kovarianz	116,211	188,466
	N	29	29

**. Die Korrelation ist auf dem Niveau von 0,01 (2-seitig) signifikant.

Fig. 57: Correlation between progresses of the concept "Dezentralisierung" extracted from CW_{5k} and from CW_{5kb} corpus

From this purely statistical perspective a high similarity between both test sets is indicated. The analysis now focuses on the semantic perspective with the objective of comparing the knowledge extracted from CW_{5k} with CW_{5kb} from a domain-expert point of view.

Table 40 documents the first ten leading aggregated concepts within corpus segments and drill down to term level in CW_{5kb}. The concepts found will be compared with those documented in Table 23. This is a limited list and only for an introduction of results. For a complete list refer to Appendix (see Table

73). The leading concepts are shown separately for each corpus segment C_C and C_V[34].

Judged from a domain-expert perspective (without calculation of support) on aggregated concept level compared to CW_{5k} no matching results were found for both C_C and C_V segments. The order of the concepts was found to be completely different for CW_{5kb}.

The dates of from-to periods (TermFirstOcc and TermLastOcc measures) for term “Chipcom” match exactly.

The exemplarily analysis of the “Vendor” dimension resulted in a curious result for the C_C segment: Where the term “IBM” matches as leading within all three time segments, the term “Microsoft” appears within segment C_C of CW_{5kb}. According to the evaluation results from CW_{5k} this term should belong to C_V instead of C_C. Another example is the vendor company “Oracle” which was extracted as leading within C_V for 1975 time segment, but Oracle was founded in 1977 and therefore did not exist beforehand.

These examples show the bias effect of non-target data within the analysed corpus. In this example a surrounding advertisement term was extracted as a leading concept due to the pure quantitative approach of extraction using TRQ threshold measures.

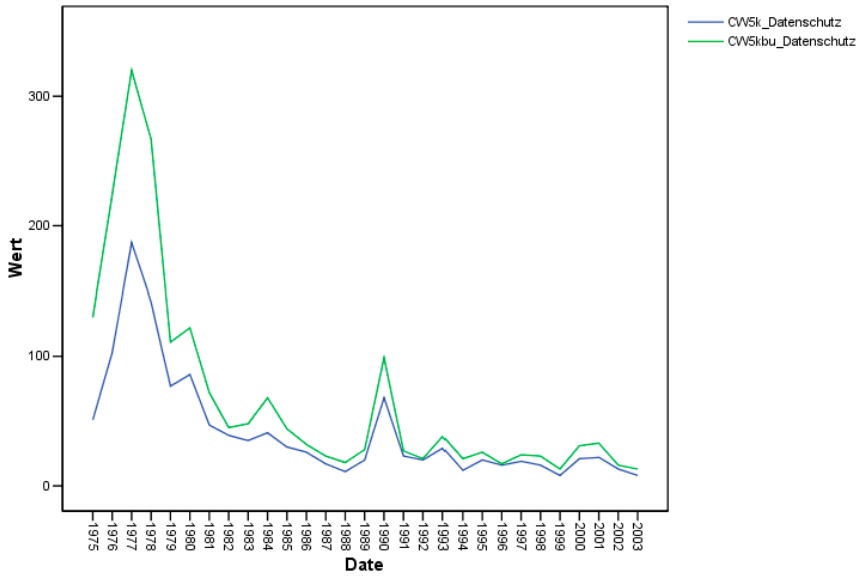
[34] Even if the names of concepts are equal within C_C and C_V, the terms assigned to these concepts are not the same, because the terms were automatically assigned to one of both corpus segments depending on their persistence in time.

Table 40: CW_{5kb}, first ten leading aggregated the concepts within corpus segments and drill-down to term level

	CW_{5kb}		
Date	1975	1988	2003
Cc_CountThresU_Dim	IT	IT	IT
	Science	Vendor	Currency
	Currency	Science	Profession
	Vendor	SocialFramework	Vendor
	Business	Currency	Science
	Profession	Profession	Economy
	Economy	Economy	Event
	SocialFramework	Business	Business
	Geography	Performance	Geography
	Customer	Name	
Cv_CountThresU_Dim	Event	Vendor	Currency
	OS	ProgLanguage	Norm
	Name	Profession	OS
	Vendor	Name	Profession
	Institute	OS	Customer
	Business	ITProduct	ITProduct
	Geography	IT	Vendor
	Economy	Institute	ProgLanguage
	IT	Business	IT
	Currency	Norm	Economy
Chipcom.TermFirstOcc	1987		
Chipcom.TermLastOcc	1998		
Cc_CountThresU_Vendor	IBM	IBM	IBM
	Microsoft	Microsoft	Siemens
	Siemens	Fujitsu	Microsoft
	HP	DEC	HP
	Apple	Apple	Sharp
	Nixdorf	Digital	SAP
		SAP	Intel
		Siemens	
		HP	
Cv_CountThresU_Vendor	Telekom	Telekom	Novell
	3Com	Bertelsmann	Oracle
	Vodafone	Novell	SCO
	Infineon	SAS	Vodafone
	Oracle	Hyperion	Borland
	Ariba	Infineon	Fujitsu-Siemens
	Lycos	Toshiba	Abb
	T-Online	Mobilcom	Microsofts
	AMD	Siebel	Telekom
	Matsushita	Oracle	EDS

4.3.2.5.2 Semantic analysis of CW corpus test set CW_{5kbu}

The similarity based on simple correlation analysis for progress paths of pre-selected concepts was again found to be significant with a high correlation between selected concepts from CW_{5k} and CW_{5kbu} (see Fig. 58 to Fig. 64).

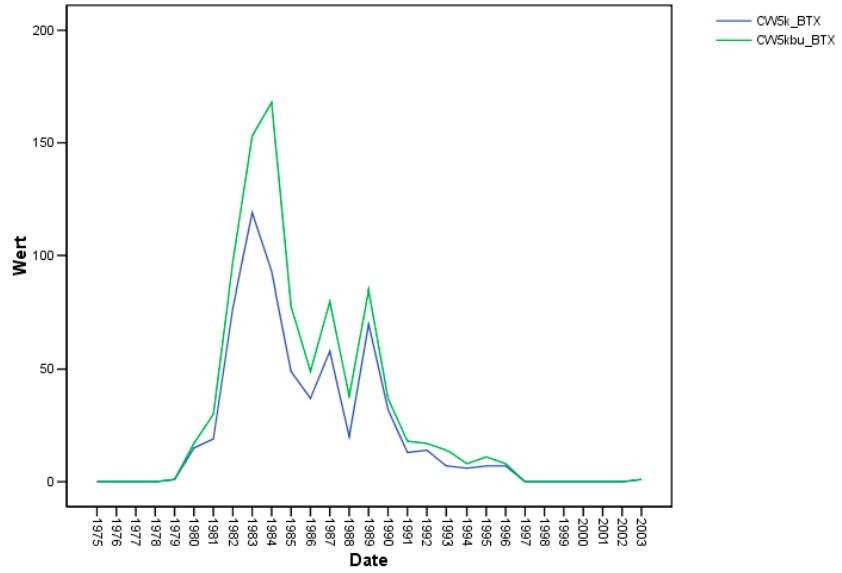


Korrelationen

		CW5k_ Datenschutz	CW5kbu_ Datenschutz
CW5k_Datenschutz	Korrelation nach Pearson	1	,980**
	Signifikanz (2-seitig)		,000
	Quadratsummen und Kreuzprodukte	49486,828	90612,172
	Kovarianz	1767,387	3236,149
	N	29	29
CW5kbu_Datenschutz	Korrelation nach Pearson	,980**	1
	Signifikanz (2-seitig)	,000	
	Quadratsummen und Kreuzprodukte	90612,172	172614,828
	Kovarianz	3236,149	6164,815
	N	29	29

**. Die Korrelation ist auf dem Niveau von 0,01 (2-seitig) signifikant.

Fig. 58: Correlation between progresses of the concept "Datenschutz" extracted from CW_{5k} and from CW_{5kbu} corpus

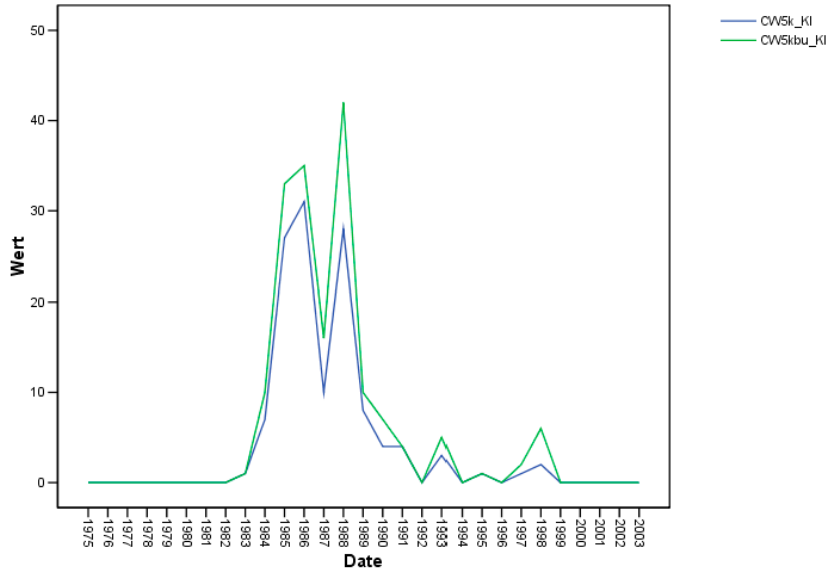


Korrelationen

		CW5k_BTX	CW5kbu_BTX
CW5k_BTX	Korrelation nach Pearson	1	,981**
	Signifikanz (2-seitig)		,000
	Quadratsummen und Kreuzprodukte	28987,310	40807,345
	Kovarianz	1035,261	1457,405
	N	29	29
CW5kbu_BTX	Korrelation nach Pearson	,981**	1
	Signifikanz (2-seitig)	,000	
	Quadratsummen und Kreuzprodukte	40807,345	59658,828
	Kovarianz	1457,405	2130,672
	N	29	29

**. Die Korrelation ist auf dem Niveau von 0,01 (2-seitig) signifikant.

Fig. 59: Correlation between progresses of the concept "BTX" extracted from CW_{5k} and from CW_{5kbu} corpus

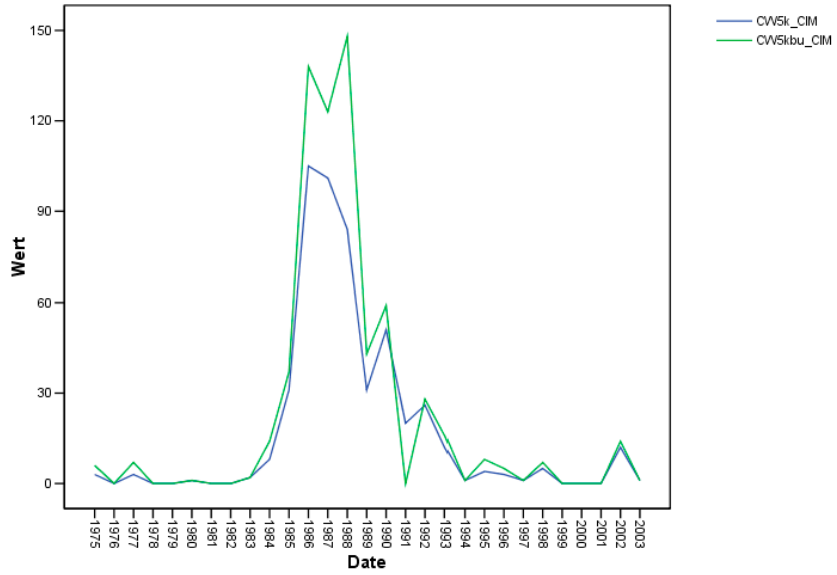


Korrelationen

		CW5k_KI	CW5kbu_KI
CW5k_KI	Korrelation nach Pearson	1	,988**
	Signifikanz (2-seitig)		,000
	Quadratsummen und Kreuzprodukte	2178,828	2783,759
	Kovarianz	77,815	99,420
	N	29	29
CW5kbu_KI	Korrelation nach Pearson	,988**	1
	Signifikanz (2-seitig)	,000	
	Quadratsummen und Kreuzprodukte	2783,759	3645,862
	Kovarianz	99,420	130,209
	N	29	29

**. Die Korrelation ist auf dem Niveau von 0,01 (2-seitig) signifikant.

Fig. 60: Correlation between progresses of the concept "KI" extracted from CW_{5k} and from CW_{5kbu} corpus

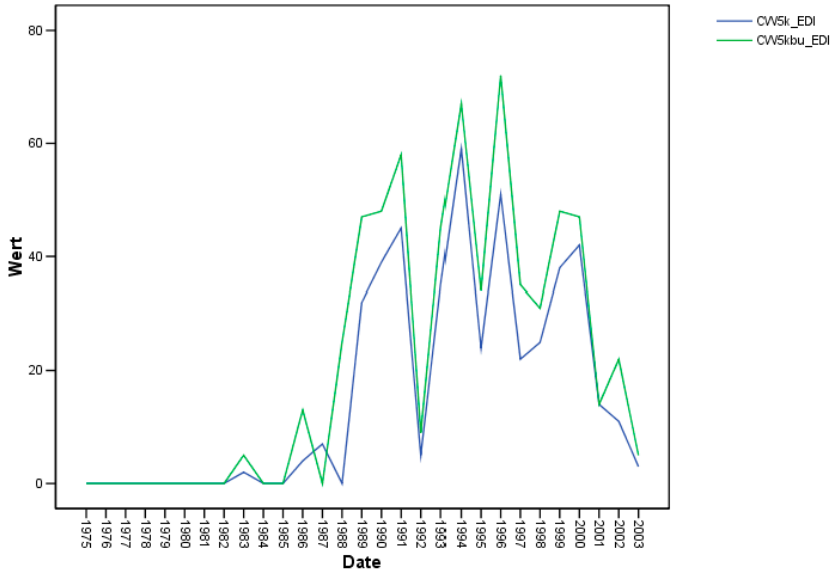


Korrelationen

		CW5k_CIM	CW5kbu_CIM
CW5k_CIM	Korrelation nach Pearson	1	,976**
	Signifikanz (2-seitig)		,000
	Quadratsummen und Kreuzprodukte	25515,034	34687,310
	Kovarianz	911,251	1238,833
	N	29	29
CW5kbu_CIM	Korrelation nach Pearson	,976**	1
	Signifikanz (2-seitig)	,000	
	Quadratsummen und Kreuzprodukte	34687,310	49463,793
	Kovarianz	1238,833	1766,564
	N	29	29

**. Die Korrelation ist auf dem Niveau von 0,01 (2-seitig) signifikant.

Fig. 61: Correlation between progresses of the concept "CIM" extracted from CW_{5k} and from CW_{5kbu} corpus

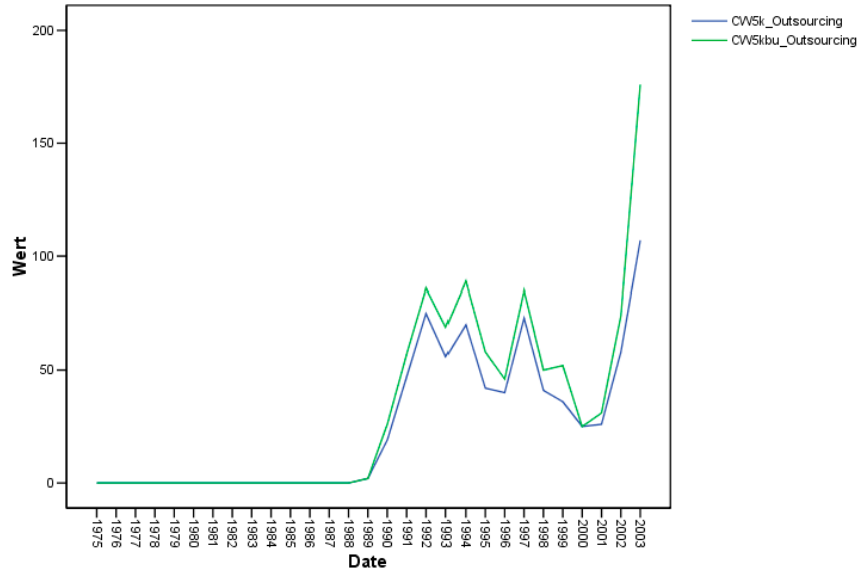


Korrelationen

		CW5k_EDI	CW5kbu_EDI
CW5k_EDI	Korrelation nach Pearson	1	,967**
	Signifikanz (2-seitig)		,000
	Quadratsummen und Kreuzprodukte	9956,759	12034,310
	Kovarianz	355,599	429,797
	N	29	29
CW5kbu_EDI	Korrelation nach Pearson	,967**	1
	Signifikanz (2-seitig)	,000	
	Quadratsummen und Kreuzprodukte	12034,310	15565,172
	Kovarianz	429,797	555,899
	N	29	29

**. Die Korrelation ist auf dem Niveau von 0,01 (2-seitig) signifikant.

Fig. 62: Correlation between progresses of the concept "EDI" extracted from CW_{5k} and from CW_{5kbu} corpus

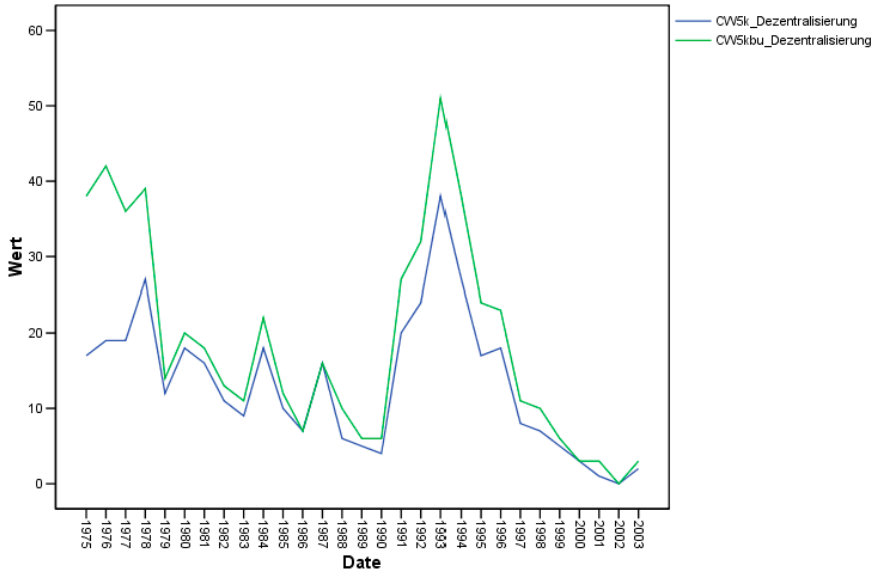


Korrelationen

		CW5k_ Outsourcing	CW5kbu_ Outsourcing
CW5k_Outsourcing	Korrelation nach Pearson	1	,983**
	Signifikanz (2-seitig)		,000
	Quadratsummen und Kreuzprodukte	26291,793	35784,448
	Kovarianz	938,993	1278,016
	N	29	29
CW5kbu_Outsourcing	Korrelation nach Pearson	,983**	1
	Signifikanz (2-seitig)	,000	
	Quadratsummen und Kreuzprodukte	35784,448	50385,862
	Kovarianz	1278,016	1799,495
	N	29	29

**. Die Korrelation ist auf dem Niveau von 0,01 (2-seitig) signifikant.

Fig. 63: Correlation between progresses of the concept "Outsourcing" extracted from CW_{5k} and from CW_{5kbu} corpus



Korrelationen

		CW5k_ Dezentralisierung	CW5kbu_ Dezentralisierung
CW5k_Dezentralisierung	Korrelation nach Pearson	1	,936**
	Signifikanz (2-seitig)		,000
	Quadratsummen und Kreuzprodukte	2301,310	3310,414
	Kovarianz	82,190	118,229
	N	29	29
CW5kbu_ Dezentralisierung	Korrelation nach Pearson	,936**	1
	Signifikanz (2-seitig)	,000	
	Quadratsummen und Kreuzprodukte	3310,414	5434,552
	Kovarianz	118,229	194,091
	N	29	29

**. Die Korrelation ist auf dem Niveau von 0,01 (2-seitig) signifikant.

Fig. 64: Correlation between progresses of the concept "Dezentralisierung" extracted from CW_{5k} and from CW_{5kbu} corpus

Table 41 documents the first ten leading aggregated concepts within corpus segments and drill-down to term level in CW_{5kbu}. The concepts found will be compared with those documented in Table 23. For the complete list refer to Appendix (see Table 74).

Quite similar to the results derived from CW_{5kb}, on aggregated concept level compared to CW_{5k} no matching results were found for both, C_C and C_V segments. The order of the concepts for CW_{5kbu} was found to be completely different.

The dates of from-to periods (TermFirstOcc and TermLastOcc measures) for the term “Chipcom” matched exactly.

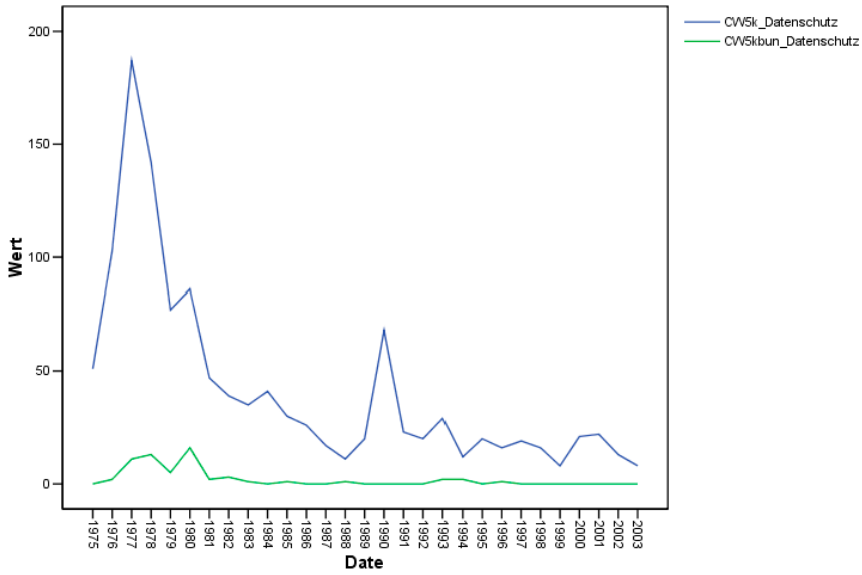
The exemplarily analysis of the “Vendor” dimension again led to curious results for the C_C segment: Where the term “IBM” matches as leading within all three time segments, the term “Microsoft” appears within segment C_C of CW_{5kb}. According to the evaluation results from CW_{5k} this term should belong to C_V instead of C_C. Another example is the vendor company “Oracle” that was extracted as one leading term within C_V for the 1975 time segment, but Oracle was founded in 1977 and therefore did not exist beforehand. The biasing effect of surrounding advertisement found for CW_{5kb} was also present for CW_{5kb}.

Table 41: CW_{5kbu}, first ten leading aggregated the concepts within corpus segments and drill-down to term level

	CW5kbu		
Date	1975	1988	2003
Cc_CountThresU_Dim	IT	IT	IT
	Economy	Economy	Vendor
	Vendor	Vendor	Event
	Business	Business	Economy
	Geography	Science	Business
Cv_CountThresU_Dim	Event	Event	Currency
	OS	OS	Norm
	Vendor	Vendor	OS
	Name	Name	Profession
	Institute	Institute	ITProduct
	Business	Business	Vendor
	Geography	Geography	Customer
	Economy	Economy	ProgLanguage
	IT	IT	IT
	Currency	Currency	Economy
Chipcom.TermFirstOcc	1987		
Chipcom.TermLastOcc	1998		
Cc_CountThresU_Vendor	IBM	IBM	IBM
		Microsoft	
Cv_CountThresU_Vendor	Telekom	Telekom	Novell
	3Com	Bertelsmann	Oracle
	Vodafone	Novell	SCO
	Infineon	SAS	Vodafone
	Oracle	Hyperion	Borland
	Ariba	Infineon	Fujitsu-Siemens
	Lycos	Toshiba	Microsofts
	T-Online	Mobilcom	Abb
	AMD	Siebel	Telekom
	Matsushita	Oracle	EDS

4.3.2.5.3 Semantic analysis of CW corpus test set CW_{5kbun} and CW_{5kbun2}

CW_{5kbun} and CW_{5kbun2} represent type “bn” corpora within this analysis. The results of correlation analyses for progress paths of pre-selected concepts from CW_{5k} and CW_{5kbun} and CW_{5kbun2} are shown in Fig. 66 to Fig. 85.

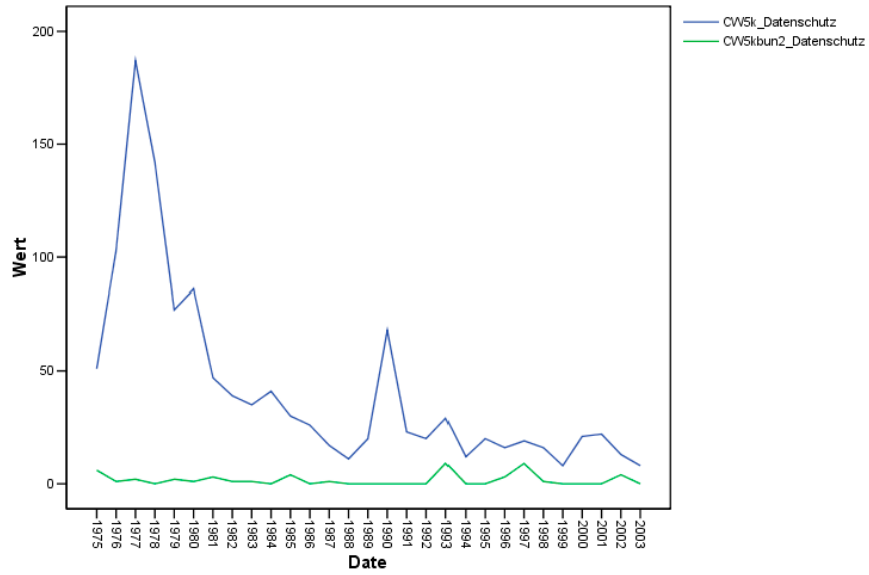


Korrelationen

		CW5k_ Datenschutz	CW5kbun_ Datenschutz
CW5k_Datenschutz	Korrelation nach Pearson	1	,774**
	Signifikanz (2-seitig)		,000
	Quadratsummen und Kreuzprodukte	49486,828	3757,759
	Kovarianz	1767,387	134,206
	N	29	29
CW5kbun_Datenschutz	Korrelation nach Pearson	,774**	1
	Signifikanz (2-seitig)	,000	
	Quadratsummen und Kreuzprodukte	3757,759	475,862
	Kovarianz	134,206	16,995
	N	29	29

**. Die Korrelation ist auf dem Niveau von 0,01 (2-seitig) signifikant.

Fig. 65: Correlation between progresses of the concept "Datenschutz" extracted from CW_{5k} and from CW_{5kbun} test set



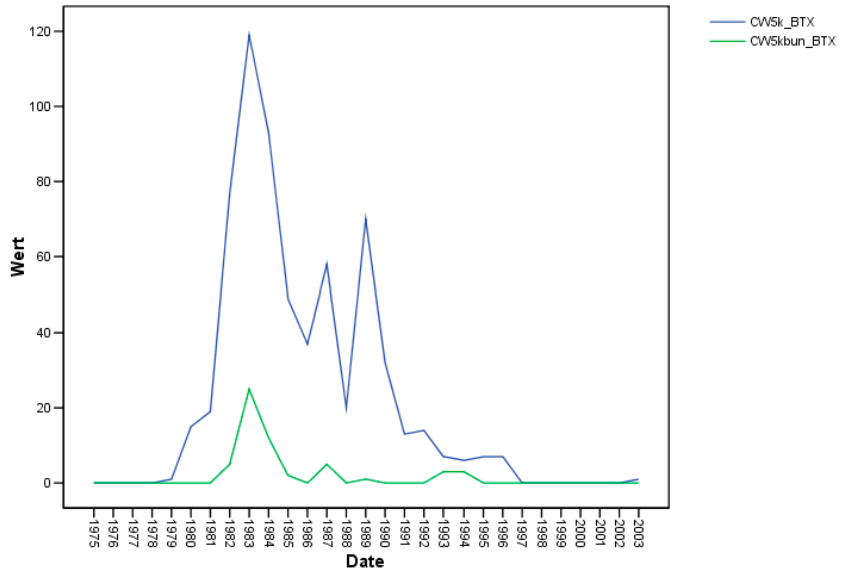
Korrelationen

		CW5k_ Datenschutz	CW5kbun2_ Datenschutz
CW5k_Datenschutz	Korrelation nach Pearson	1	-,025
	Signifikanz (2-seitig)		,898
	Quadratsummen und Kreuzprodukte	49486,828	-74,793
	Kovarianz	1767,387	-2,671
	N	29	29
CW5kbun2_Datenschutz	Korrelation nach Pearson	-,025	1
	Signifikanz (2-seitig)	,898	
	Quadratsummen und Kreuzprodukte	-74,793	182,552
	Kovarianz	-2,671	6,520
	N	29	29

Fig. 66: Correlation between progresses of the concept "Datenschutz" extracted from CW_{5k} and from CW_{5kbun2} test set

The equality between the CW5k progress path for the concept “Datenschutz” and that extracted from the test sets in focus here was found to be significant for test set CW_{5kbun} and not present for test set CW_{5kbun2}.

For the concept “BTX” a significant correlation was found in both cases, whereas the concept “KI” was not present within CW_{5kbun}, but led to a significant positive correlation with the result from CW_{5k} with test set CW_{5kbun2}.

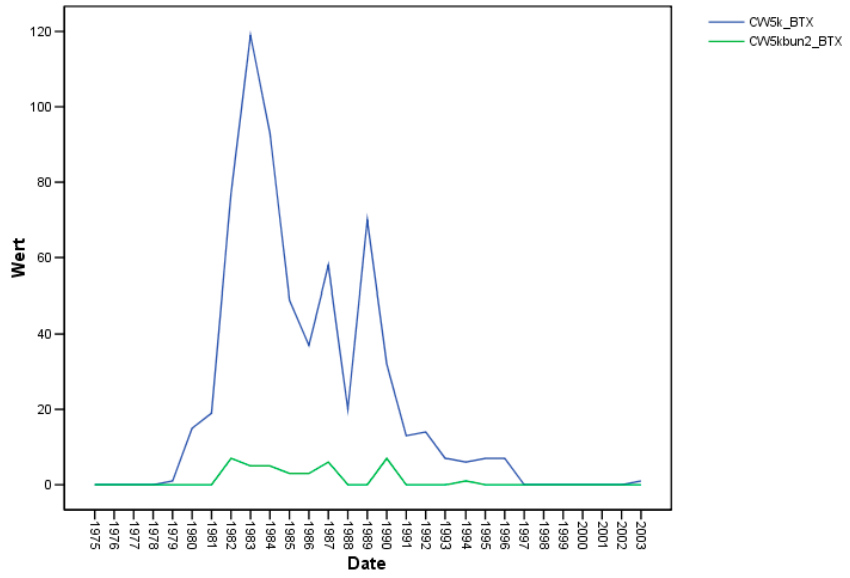


Korrelationen

		CW5k_BTX	CW5kbun_BTX
CW5k_BTX	Korrelation nach Pearson	1	,808**
	Signifikanz (2-seitig)		,000
	Quadratsummen und Kreuzprodukte	28987,310	3727,483
	Kovarianz	1035,261	133,124
	N	29	29
CW5kbun_BTX	Korrelation nach Pearson	,808**	1
	Signifikanz (2-seitig)	,000	
	Quadratsummen und Kreuzprodukte	3727,483	733,862
	Kovarianz	133,124	26,209
	N	29	29

**. Die Korrelation ist auf dem Niveau von 0,01 (2-seitig) signifikant.

Fig. 67: Correlation between progresses of the concept "BTX" extracted from CW_{5k} and from CW_{5kbun} test set

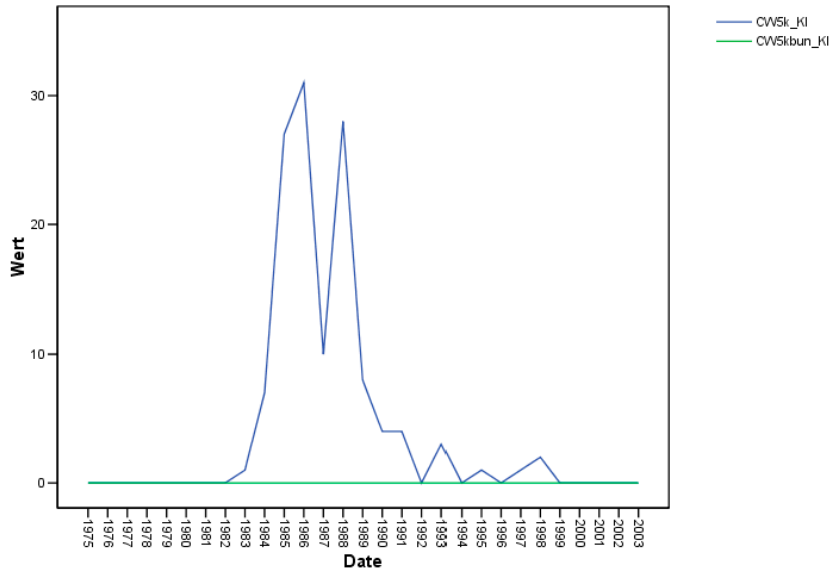


Korrelationen

		CW5k_BTX	CW5kbun2_BTX
CW5k_BTX	Korrelation nach Pearson	1	,759**
	Signifikanz (2-seitig)		,000
	Quadratsummen und Kreuzprodukte	28987,310	1612,069
	Kovarianz	1035,261	57,574
	N	29	29
CW5kbun2_BTX	Korrelation nach Pearson	,759**	1
	Signifikanz (2-seitig)	,000	
	Quadratsummen und Kreuzprodukte	1612,069	155,793
	Kovarianz	57,574	5,564
	N	29	29

**. Die Korrelation ist auf dem Niveau von 0,01 (2-seitig) signifikant.

Fig. 68: Correlation between progresses of the concept "BTX" extracted from CW_{5k} and from CW_{5kbun2} test set

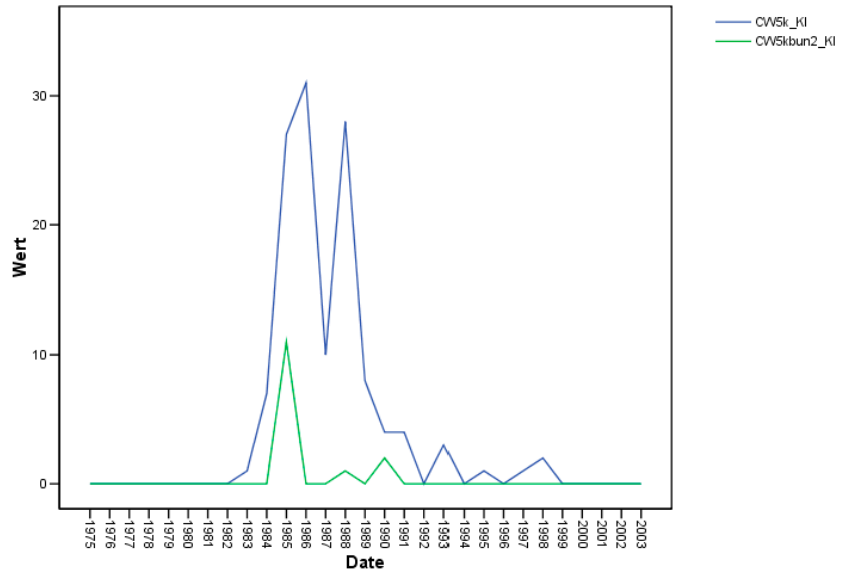


Korrelationen

		CW5k_KI	CW5kbun_KI
CW5k_KI	Korrelation nach Pearson	1	.[a]
	Signifikanz (2-seitig)		.
	Quadratsummen und Kreuzprodukte	2178,828	,000
	Kovarianz	77,815	,000
	N	29	29
CW5kbun_KI	Korrelation nach Pearson	.[a]	.[a]
	Signifikanz (2-seitig)	.	
	Quadratsummen und Kreuzprodukte	,000	,000
	Kovarianz	,000	,000
	N	29	29

a. Kann nicht berechnet werden, da mindestens eine der Variablen konstant ist.

Fig. 69: Correlation between progresses of the concept "KI" extracted from CW_{5k} and from CW_{5kbun} test set

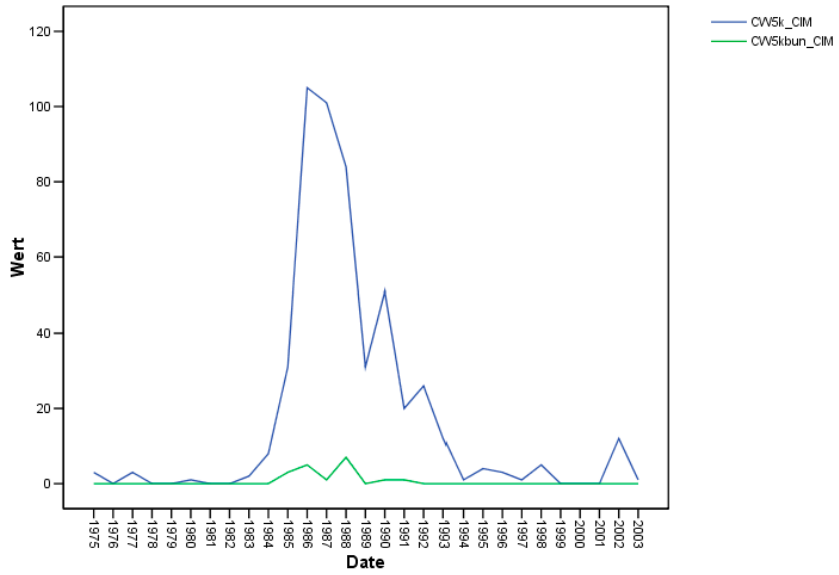


Korrelationen

		CW5k_KI	CW5kbun2_KI
CW5k_KI	Korrelation nach Pearson	1	,533**
	Signifikanz (2-seitig)		,003
	Quadratsummen und Kreuzprodukte	2178,828	271,690
	Kovarianz	77,815	9,703
	N	29	29
CW5kbun2_KI	Korrelation nach Pearson	,533**	1
	Signifikanz (2-seitig)	,003	
	Quadratsummen und Kreuzprodukte	271,690	119,241
	Kovarianz	9,703	4,259
	N	29	29

**. Die Korrelation ist auf dem Niveau von 0,01 (2-seitig) signifikant.

Fig. 70: Correlation between progresses of the concept "KI" extracted from CW_{5k} and from CW_{5kbun2} test set

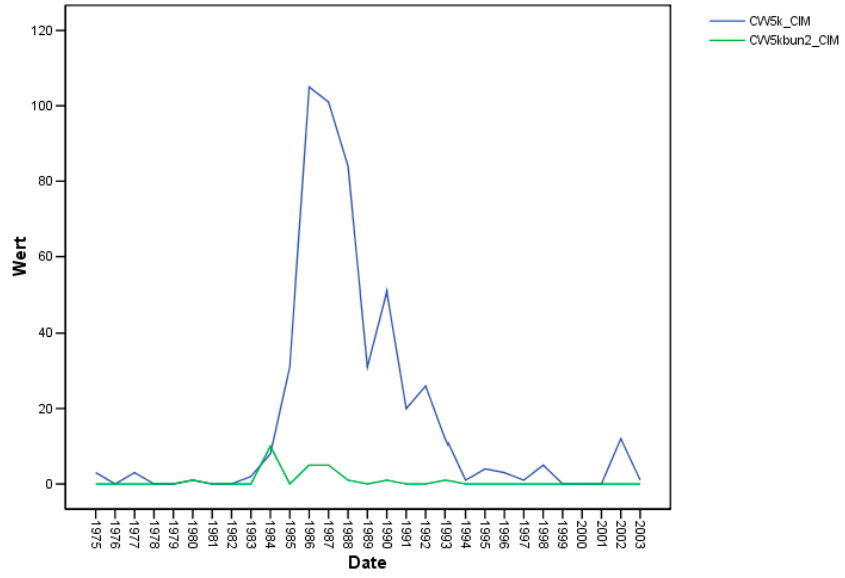


Korrelationen

		CW5k_CIM	CW5kbun_CIM
CW5k_CIM	Korrelation nach Pearson	1	,770**
	Signifikanz (2-seitig)		,000
	Quadratsummen und Kreuzprodukte	25515,034	1064,552
	Kovarianz	911,251	38,020
	N	29	29
CW5kbun_CIM	Korrelation nach Pearson	,770**	1
	Signifikanz (2-seitig)	,000	
	Quadratsummen und Kreuzprodukte	1064,552	74,828
	Kovarianz	38,020	2,672
	N	29	29

**. Die Korrelation ist auf dem Niveau von 0,01 (2-seitig) signifikant.

Fig. 71: Correlation between progresses of the concept "CIM" extracted from CW_{5k} and from CW_{5kbun} test set

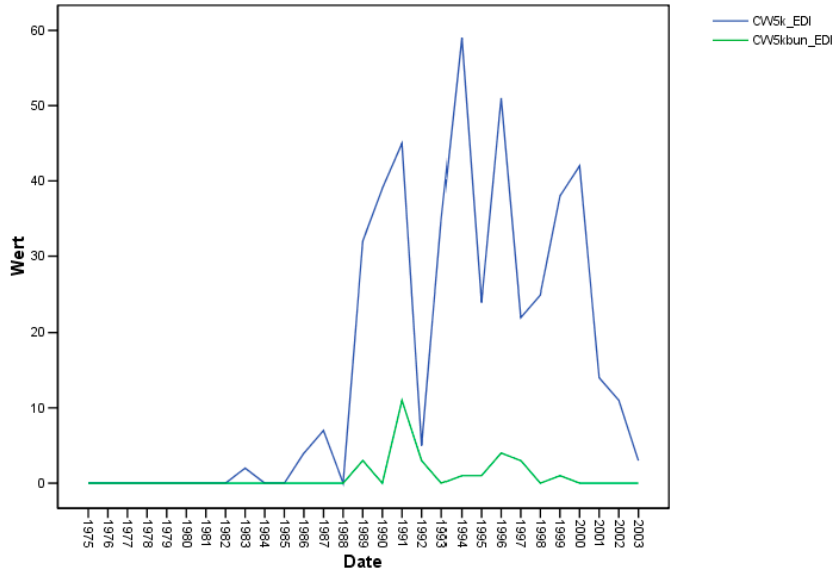


Korrelationen

		CW5k_CIM	CW5kbun2_CIM
CW5k_CIM	Korrelation nach Pearson	1	,454*
	Signifikanz (2-seitig)		,013
	Quadratsummen und Kreuzprodukte	25515,034	840,069
	Kovarianz	911,251	30,002
	N	29	29
CW5kbun2_CIM	Korrelation nach Pearson	,454*	1
	Signifikanz (2-seitig)	,013	
	Quadratsummen und Kreuzprodukte	840,069	134,138
	Kovarianz	30,002	4,791
	N	29	29

*. Die Korrelation ist auf dem Niveau von 0,05 (2-seitig) signifikant.

Fig. 72: Correlation between progresses of the concept "CIM" extracted from CW_{5k} and from CW_{5kbun2} test set

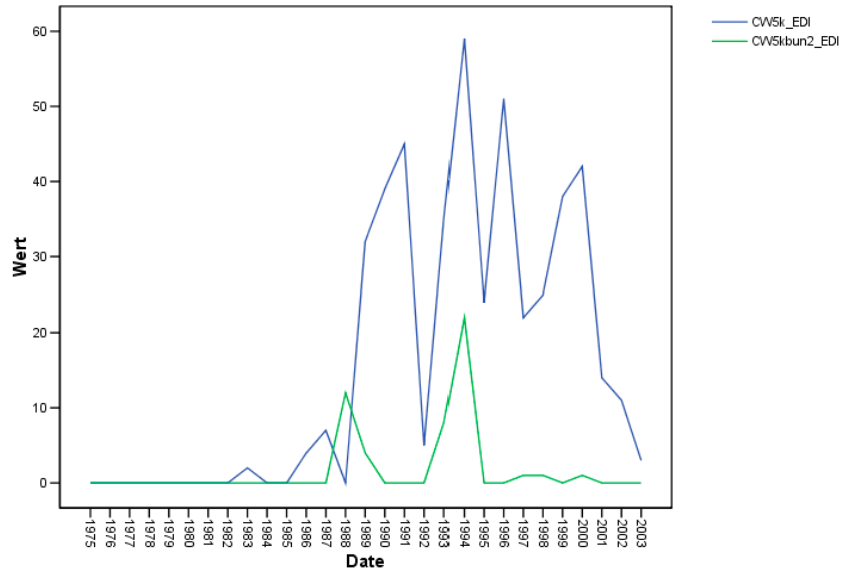


Korrelationen

		CW5k_EDI	CW5kbun_EDI
CW5k_EDI	Korrelation nach Pearson	1	,480**
	Signifikanz (2-seitig)		,008
	Quadratsummen und Kreuzprodukte	9956,759	570,586
	Kovarianz	355,599	20,378
	N	29	29
CW5kbun_EDI	Korrelation nach Pearson	,480**	1
	Signifikanz (2-seitig)	,008	
	Quadratsummen und Kreuzprodukte	570,586	141,862
	Kovarianz	20,378	5,067
	N	29	29

**. Die Korrelation ist auf dem Niveau von 0,01 (2-seitig) signifikant.

Fig. 73: Correlation between progresses of the concept "EDI" extracted from CW_{5k} and from CW_{5kbun} test set

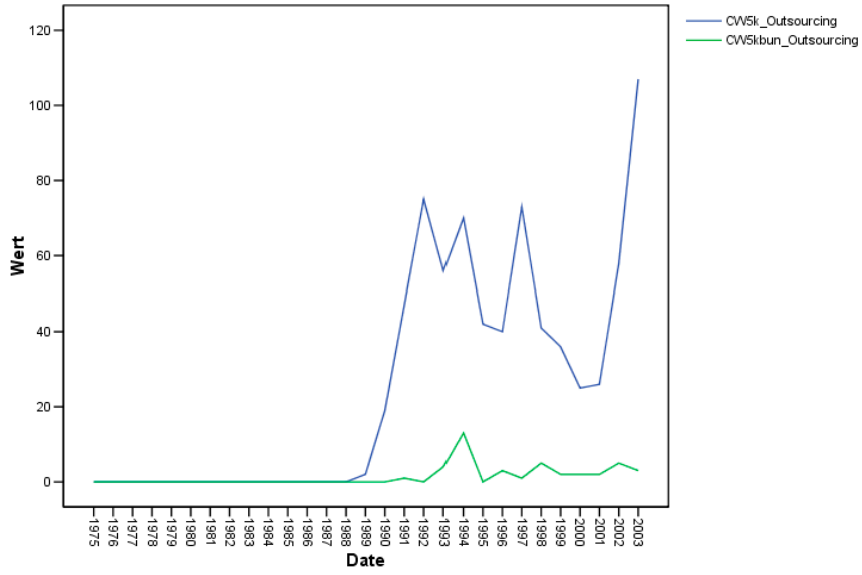


Korrelationen

		CW5k_EDI	CW5kbun2_EDI
CW5k_EDI	Korrelation nach Pearson	1	,408*
	Signifikanz (2-seitig)		,028
	Quadratsummen und Kreuzprodukte	9956,759	1021,138
	Kovarianz	355,599	36,469
	N	29	29
CW5kbun2_EDI	Korrelation nach Pearson	,408*	1
	Signifikanz (2-seitig)	,028	
	Quadratsummen und Kreuzprodukte	1021,138	628,207
	Kovarianz	36,469	22,436
	N	29	29

*. Die Korrelation ist auf dem Niveau von 0,05 (2-seitig) signifikant.

Fig. 74: Correlation between progresses of the concept "EDI" extracted from CW_{5k} and from CW_{5kbun2} test set

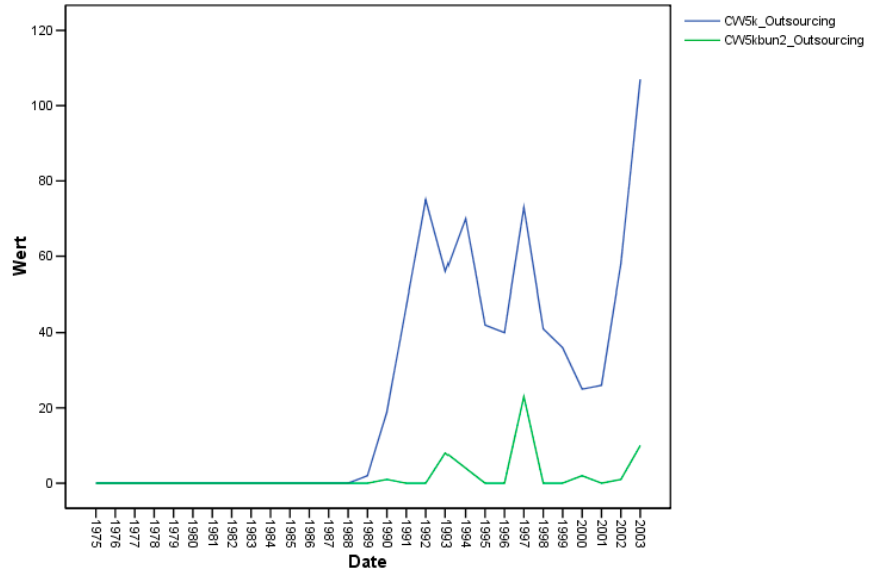


Korrelationen

		CW5k_ Outsourcing	CW5kbun_ Outsourcing
CW5k_Outsourcing	Korrelation nach Pearson	1	,576**
	Signifikanz (2-seitig)		,001
	Quadratsummen und Kreuzprodukte	26291,793	1350,310
	Kovarianz	938,993	48,225
	N	29	29
CW5kbun_Outsourcing	Korrelation nach Pearson	,576**	1
	Signifikanz (2-seitig)	,001	
	Quadratsummen und Kreuzprodukte	1350,310	209,034
	Kovarianz	48,225	7,466
	N	29	29

**. Die Korrelation ist auf dem Niveau von 0,01 (2-seitig) signifikant.

Fig. 75: Correlation between progresses of the concept "Outsourcing" extracted from CW_{5k} and from CW_{5kbun} test set



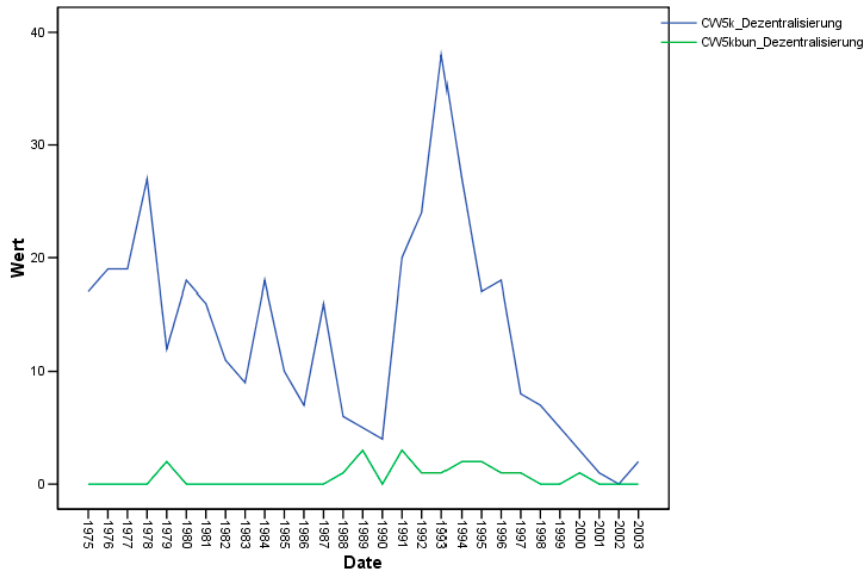
Korrelationen

		CW5k_ Outsourcing	CW5kbun2_ Outsourcing
CW5k_Outsourcing	Korrelation nach Pearson	1	,587**
	Signifikanz (2-seitig)		,001
	Quadratsummen und Kreuzprodukte	26291,793	2392,517
	Kovarianz	938,993	85,447
	N	29	29
CW5kbun2_Outsourcing	Korrelation nach Pearson	,587**	1
	Signifikanz (2-seitig)	,001	
	Quadratsummen und Kreuzprodukte	2392,517	632,207
	Kovarianz	85,447	22,579
	N	29	29

**. Die Korrelation ist auf dem Niveau von 0,01 (2-seitig) signifikant.

Fig. 76: Correlation between progresses of the concept "Outsourcing" extracted from CW_{5k} and from CW_{5kbun2} test set

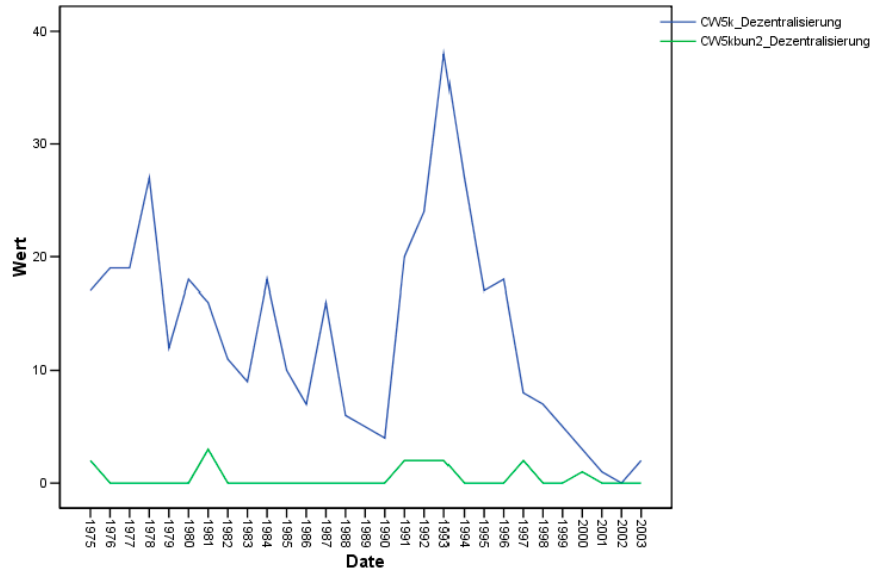
For the concepts “CIM”, “EDI” and “Outsourcing” significant correlations were found.



Korrelationen

		CW5k_ Dezentralisierung	CW5kbun _Dezentralisierung
CW5k_Dezentralisierung	Korrelation nach Pearson	1	,191
	Signifikanz (2-seitig)		,321
	Quadratsummen und Kreuzprodukte	2301,310	45,655
	Kovarianz	82,190	1,631
	N	29	29
CW5kbun_ Dezentralisierung	Korrelation nach Pearson	,191	1
	Signifikanz (2-seitig)	,321	
	Quadratsummen und Kreuzprodukte	45,655	24,828
	Kovarianz	1,631	,887
	N	29	29

Fig. 77: Correlation between progresses of the concept "Dezentralisierung" extracted from CW_{5k} and from CW_{5kbun} test set



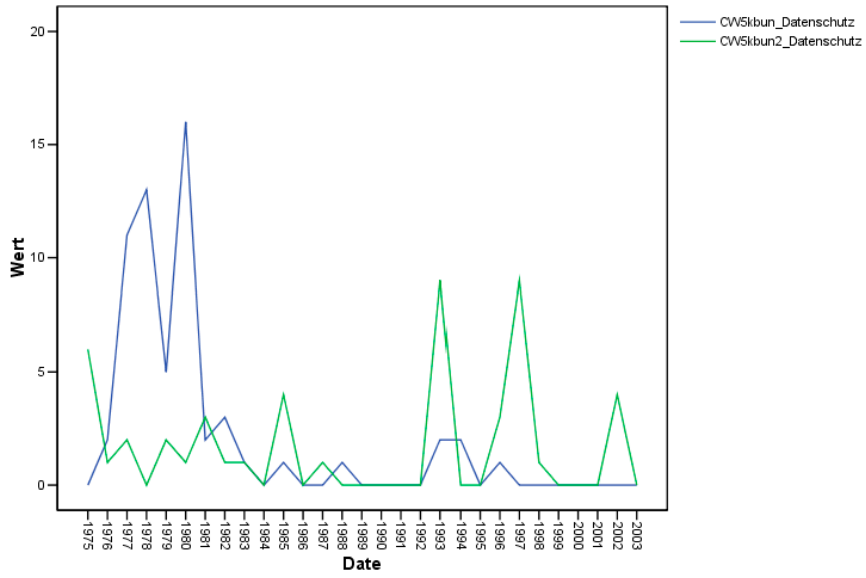
Korrelationen

		CW5k_ Dezentrali sierung	CW5kbun2 _Dezentrali sierung
CW5k_Dezentralisierung	Korrelation nach Pearson	1	,344
	Signifikanz (2-seitig)		,067
	Quadratsummen und Kreuzprodukte	2301,310	79,621
	Kovarianz	82,190	2,844
	N	29	29
CW5kbun2_ Dezentralisierung	Korrelation nach Pearson	,344	1
	Signifikanz (2-seitig)	,067	
	Quadratsummen und Kreuzprodukte	79,621	23,241
	Kovarianz	2,844	,830
	N	29	29

Fig. 78: Correlation between progresses of the concept "Dezentralisierung" extracted from CW_{5k} and from CW_{5kbun2} test set

The correlations for the concept “Dezentralisierung” are very weak in both cases.

With both these test sets built using the same corpus basis, measuring inter-test-set correlations is possible. It is to be expected that progress paths of certain pre-selected concepts (e.g., the “Dim_Mertens” taxonomy) are similar when extracted from CW_{5kbun} and CW_{5kbun2}. The results can then be interpreted as indicators for the robustness of the applied knowledge-extraction method, when applied to these test sets and their semantic similarity as well. The results of this analysis can be seen in Fig. 79 to Fig. 85.

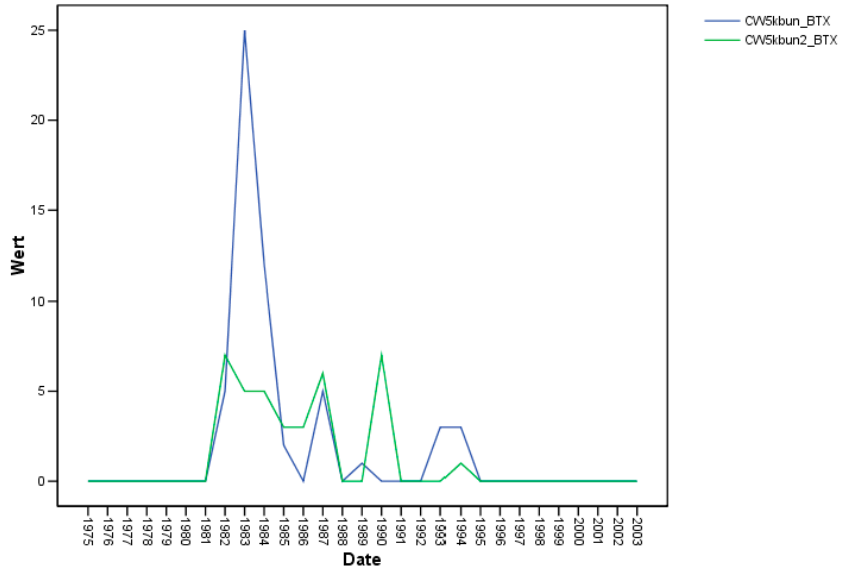


Korrelationen

		CW5kbun_ Datenschutz	CW5kbun2_ Datenschutz
CW5kbun_Datenschutz	Korrelation nach Pearson	1	-,049
	Signifikanz (2-seitig)		,803
	Quadratsummen und Kreuzprodukte	475,862	-14,310
	Kovarianz	16,995	-,511
	N	29	29
CW5kbun2_Datenschutz	Korrelation nach Pearson	-,049	1
	Signifikanz (2-seitig)	,803	
	Quadratsummen und Kreuzprodukte	-14,310	182,552
	Kovarianz	-,511	6,520
	N	29	29

Fig. 79: Correlation between progresses of the concept "Datenschutz" extracted from CW_{5kbun} and from CW_{5kbun2} corpus

The progress path for the concept “Datenschutz” was extracted very differently from both test sets with only weak correlation (see Fig. 79).



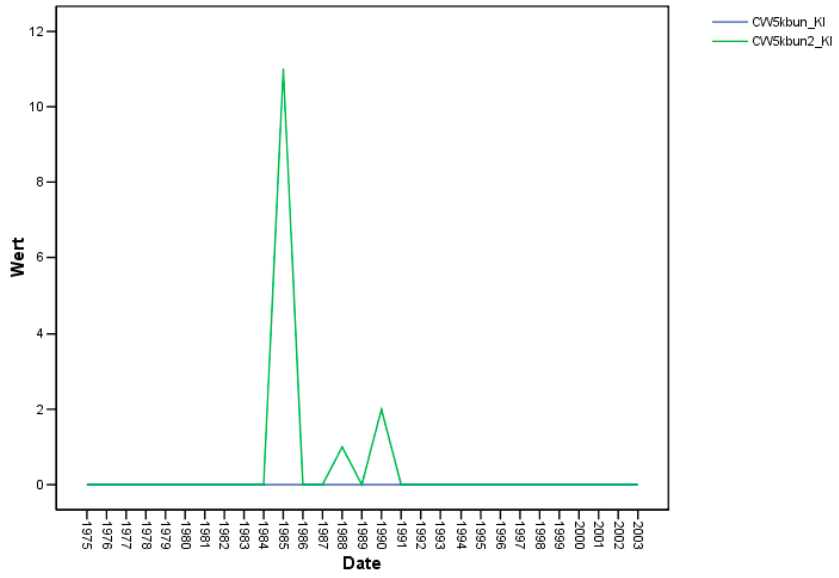
Korrelationen

		CW5kbun_ BTX	CW5kbun2_ BTX
CW5kbun_BTX	Korrelation nach Pearson	1	,555**
	Signifikanz (2-seitig)		,002
	Quadratsummen und Kreuzprodukte	733,862	187,552
	Kovarianz	26,209	6,698
	N	29	29
CW5kbun2_BTX	Korrelation nach Pearson	,555**	1
	Signifikanz (2-seitig)	,002	
	Quadratsummen und Kreuzprodukte	187,552	155,793
	Kovarianz	6,698	5,564
	N	29	29

**. Die Korrelation ist auf dem Niveau von 0,01 (2-seitig) signifikant.

Fig. 80: Correlation between progresses of the concept "BTX" extracted from CW_{5kbun} and from CW_{5kbun2} corpus

A significant positive correlation was found for the concept “BTX” (see Fig. 80).



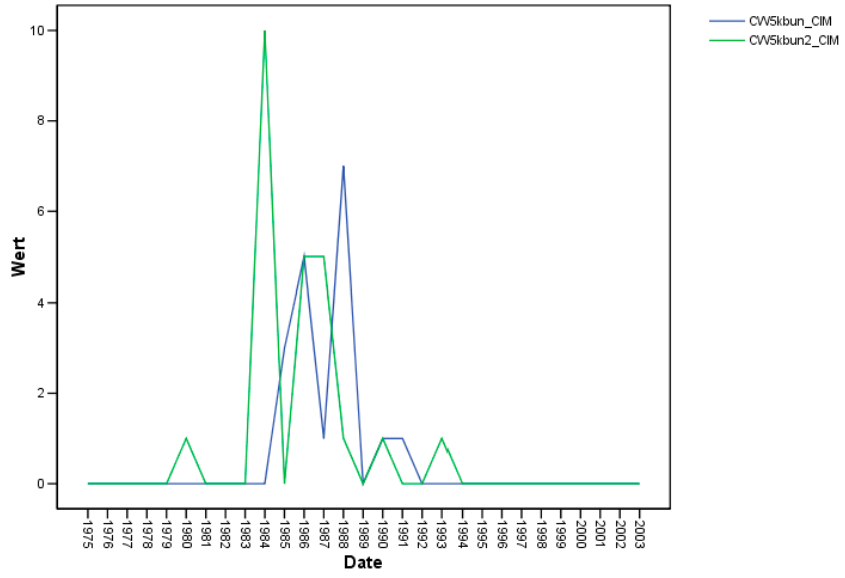
Korrelationen

		CW5kbun_KI	CW5kbun2_KI
CW5kbun_KI	Korrelation nach Pearson	.[a]	.[a]
	Signifikanz (2-seitig)		.
	Quadratsummen und Kreuzprodukte	,000	,000
	Kovarianz	,000	,000
	N	29	29
CW5kbun2_KI	Korrelation nach Pearson	.[a]	1
	Signifikanz (2-seitig)	.	
	Quadratsummen und Kreuzprodukte	,000	119,241
	Kovarianz	,000	4,259
	N	29	29

a. Kann nicht berechnet werden, da mindestens eine der Variablen konstant ist.

Fig. 81: Correlation between progresses of the concept "KI" extracted from CW_{5kbun} and from CW_{5kbun2} corpus

The concept “KI” was not present in test set CW_{5Kbun}; therefore, no correlation was found in progress paths (see Fig. 81).

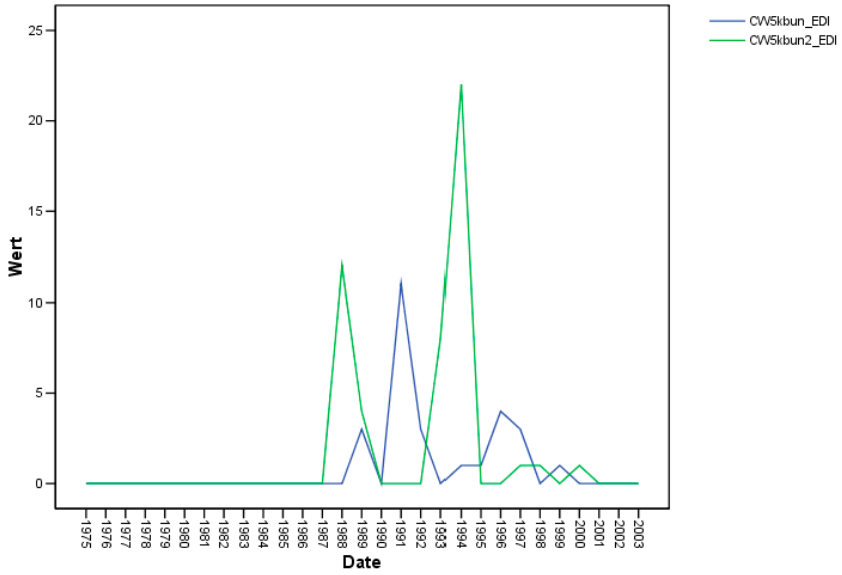


Korrelationen

		CW5kbun_CIM	CW5kbun2_CIM
CW5kbun_CIM	Korrelation nach Pearson	1	,231
	Signifikanz (2-seitig)		,229
	Quadratsummen und Kreuzprodukte	74,828	23,103
	Kovarianz	2,672	,825
	N	29	29
CW5kbun2_CIM	Korrelation nach Pearson	,231	1
	Signifikanz (2-seitig)	,229	
	Quadratsummen und Kreuzprodukte	23,103	134,138
	Kovarianz	,825	4,791
	N	29	29

Fig. 82: Correlation between progresses of the concept "CIM" extracted from CW_{5kbun} and from CW_{5kbun2} corpus

The correlation for the concept “CIM” was positive but weak (see Fig. 82).

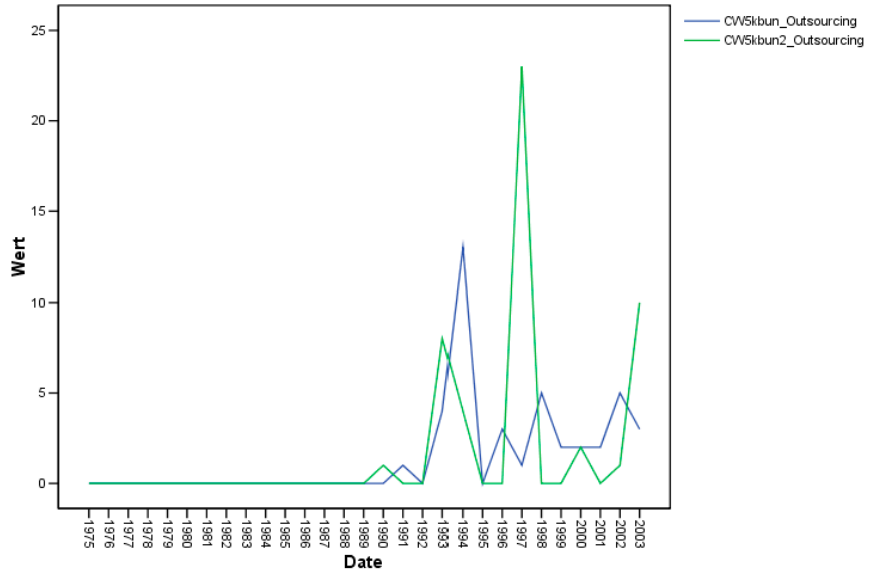


Korrelationen

		CW5kbun_EDI	CW5kbun2_EDI
CW5kbun_EDI	Korrelation nach Pearson	1	-,029
	Signifikanz (2-seitig)		,882
	Quadratsummen und Kreuzprodukte	141,862	-8,621
	Kovarianz	5,067	-,308
	N	29	29
CW5kbun2_EDI	Korrelation nach Pearson	-,029	1
	Signifikanz (2-seitig)	,882	
	Quadratsummen und Kreuzprodukte	-8,621	628,207
	Kovarianz	-,308	22,436
	N	29	29

Fig. 83: Correlation between progresses of the concept "EDI" extracted from CW_{5kbun} and from CW_{5kbun2} corpus

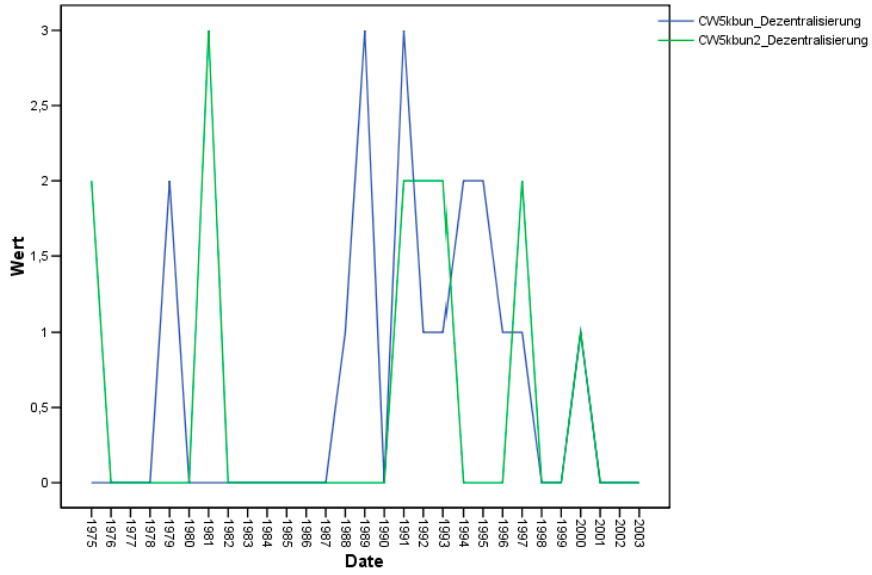
No similar progress paths were found for the concept “EDI” (see Fig. 83).



Korrelationen

		CW5kbun_ Outsourcing	CW5kbun2_ Outsourcing
CW5kbun_Outsourcing	Korrelation nach Pearson	1	,211
	Signifikanz (2-seitig)		,272
	Quadratsummen und Kreuzprodukte	209,034	76,724
	Kovarianz	7,466	2,740
	N	29	29
CW5kbun2_Outsourcing	Korrelation nach Pearson	,211	1
	Signifikanz (2-seitig)	,272	
	Quadratsummen und Kreuzprodukte	76,724	632,207
	Kovarianz	2,740	22,579
	N	29	29

Fig. 84: Correlation between progresses of the concept "Outsourcing" extracted from CW_{5kbun} and from CW_{5kbun2} corpus



Korrelationen

		CW5kbun_Dezentralisierung	CW5kbun2_Dezentralisierung
CW5kbun_Dezentralisierung	Korrelation nach Pearson	1	,179
	Signifikanz (2-seitig)		,352
	Quadratsummen und Kreuzprodukte	24,828	4,310
	Kovarianz	,887	,154
	N	29	29
CW5kbun2_Dezentralisierung	Korrelation nach Pearson	,179	1
	Signifikanz (2-seitig)	,352	
	Quadratsummen und Kreuzprodukte	4,310	23,241
	Kovarianz	,154	,830
	N	29	29

Fig. 85: Correlation between progresses of the concept "Dezentralisierung" extracted from CW_{5kbun} and from CW_{5kbun2} corpus

The correlations for the concepts “Outsourcing” and “Dezentralisierung” were very weak and not significant (see Fig. 84 and Fig. 85).

Summarizing the correlation analysis above, a low level of similarity in progress paths was found for most of the concepts that were analysed. Even if this is not a representative result from a statistical perspective, complete matching results cannot be expected from the following semantic analysis.

See Table 75, Table 76, respectively, for complete lists of extracted results. In Table 42 the first ten leading aggregated concepts within corpus segments and a drill-down to term level are shown for both test sets.

It is to be expected that both test sets represent corpora very similar to real DM scenarios: The researcher does not have knowledge about the intensity of pre-processing, but he assumes that this level is very high (with $C_G \to 0$). Then he considers the corpus length dependency of frequency-based measures (here: TRQ) and generates yearly corpus segments with an even number of terms per year. Due to the real corpus quality being realized with $C_G \geq 0$, the corpus length dependency of the TRQ measure prejudices the results of extracted knowledge.

Very few or no concepts were found for corpus segment C_C (refer to Table 23). The occurrence period for the concept “Chipcom” was found in 1994-1995 for CW_{5kbun} and 1993-1996 for CW_{5kbun2} contrary to results, extracted from CW_{5k}. Curious results were found for the drill-down within the concept “Vendor” in the domain-specific taxonomy: Companies that were definitely founded in later periods were discovered to be leading for early observation periods. This effect is realized due to the existing high share of non-target data within these datasets. Advertisements around the CW archive articles were wrongly extracted as leading concepts.

Table 42: CW_{5kbun} and CW_{5kbun2}, leading aggregated concepts within corpus segments and drill-down to term level

	CW5kbun		
Date	1975	1988	2003
Cc_CountThresU_Dim	IT	IT	IT
Cv_CountThresU_Dim	Event	OS	Event
	Vendor	Vendor	ITProduct
	Institute	ProgLanguage	Currency
	Geography	SocialFramework	OS
	Business	Science	Vendor
	Customer	Business	Economy
	OS	ITProduct	Norm
	Name	Economy	IT
	IT	Event	Geography
Chipcom.TermFirstOcc	1994		
Chipcom.TermLastOcc	1995		
Cc_CountThresU_Vendor	-	-	-
Cv_CountThresU_Vendor	Telekom	Telekom	SCO
	3Com	Bertelsmann	HP
	Vodafone	Fujitsu	Novell
	Apple	Apple	Sun
	Ariba	Digital	Dell
	HP	DEC	Oracle
	Infineon	SAS	Vodafone
	Lycos	Dell	SAP
	Oracle	Hyperion	3Com
	T-Online	Infineon	Borland

	CW_{5kbun2}		
Date	1975	1988	2003
Cc_CountThresU_Dim	IT	IT	IT
Cv_CountThresU_Dim	Event	Performance	Event
	Currency	ProgLanguage	Norm
	Vendor	Vendor	ITProduct
	Institute	Event	Vendor
	Geography	Currency	OS
	Business	OS	Currency
	Name	Science	Economy
	OS	ITProduct	IT
	SocialFramework	SocialFramework	Customer
	IT	Business	Institute
Chipcom.TermFirstOcc	1993		
Chipcom.TermLastOcc	1996		
Cc_CountThresU_Vendor	-	-	-
Cv_CountThresU_Vendor	Telekom	Telekom	Novell
	3Com	Bertelsmann	Sun
	Vodafone	Apple	Siemens
	Siemens	DEC	HP
	Apple	Digital	Borland
	Ariba	Fujitsu	Dell
	HP	Novell	Oracle
	Infineon	Dell	SAP
	Lycos	Hyperion	SCO
	Oracle	Infineon	Fujitsu-Siemens

The corpus segments were also quantitatively normalized; the extracted knowledge is not similar to that extracted from CW_{5k}. On concept level for C_C and C_V the leading concepts do not match the evaluation test set CW_{5k} (compared to Table 23).

4.3.2.6 Semantic analysis of type b corpora summary

- *The extracted knowledge was more precise the more token of target data C_T the test set contained.*
- *Quantitative normalizing of corpus time segments led to semantic information loss, when applied to type "b" corpora with the same share of target data C_T.*
- *Concepts which were pre-selected as exemplary benchmarks from external domain-expert background knowledge were extracted with less support than from type "n" corpora.*

4.4 Evaluating the impact of language of origin

For the analysis of the dependency of source language on extracted knowledge the AI1k corpus was used, which was partly translated from English into German, using optical character recognition (OCR) software. The AI1k corpus was built up by the use of yearly management reports from German and English (see Fig. 12).

4.4.1 Language fingerprint on corpus level

The Graph of TRQ measure in both AI1k corpus test sets with lower and upper confidence interval (based on the last 10-year TRQ time series) can be seen in Fig. 86.

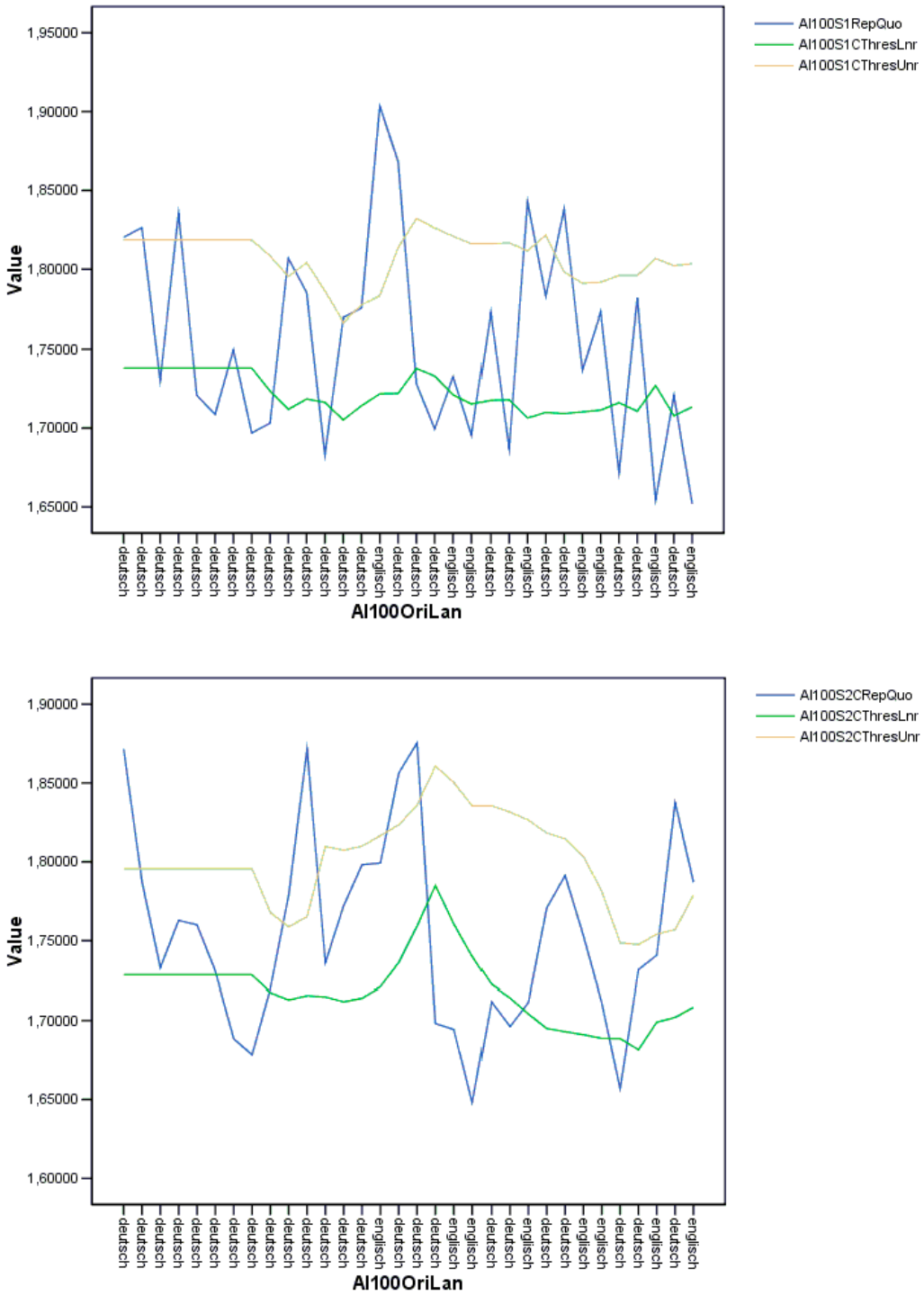
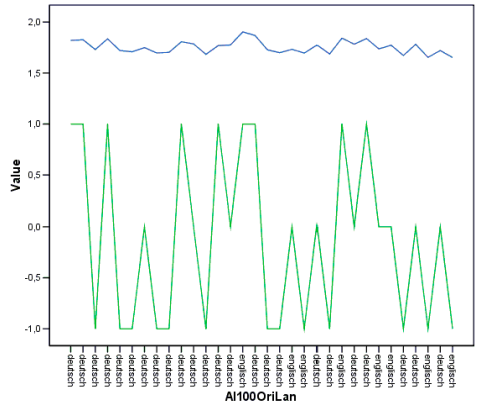


Fig. 86: Graph of TRQ measure in both AI1k corpus test sets with lower and upper confidence interval (based on last 10-year TRQ time series)

In Fig. 86 the x-axis indicates the source language: “deutsch” equals “German”, “englisch” equals “English”. It seems that the violation of the confidence interval borders is very irregular. To test the dependency on TRQ and source language it is necessary to define an indicator. The violation of the 95% confidence interval for TRQ based on a ten-year time interval was indi-

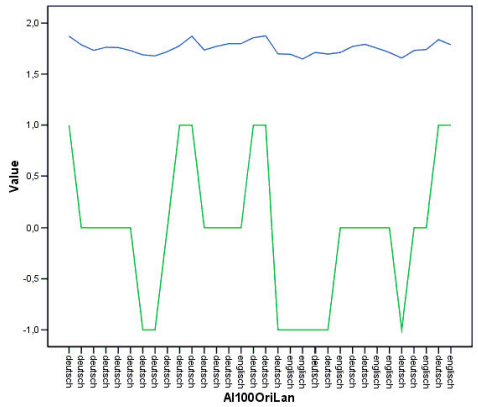
cated with an indicator ThresI for transgression (value “1”) and under run (value “-1”) of 95% confidence intervals. The default value was “0” for no violation.



Correlations

		AI100S1RepQuo	AI100S1CThresI
AI100S1RepQuo	Pearson Correlation	1	,902**
	Sig. (2-tailed)		,000
	Sum of Squares and Cross-products	,129	1,500
	Covariance	,004	,048
	N	32	32
AI100S1CThresI	Pearson Correlation	,902**	1
	Sig. (2-tailed)	,000	
	Sum of Squares and Cross-products	1,500	21,500
	Covariance	,048	,694
	N	32	32

**. Correlation is significant at the 0.01 level (2-tailed).



Correlations

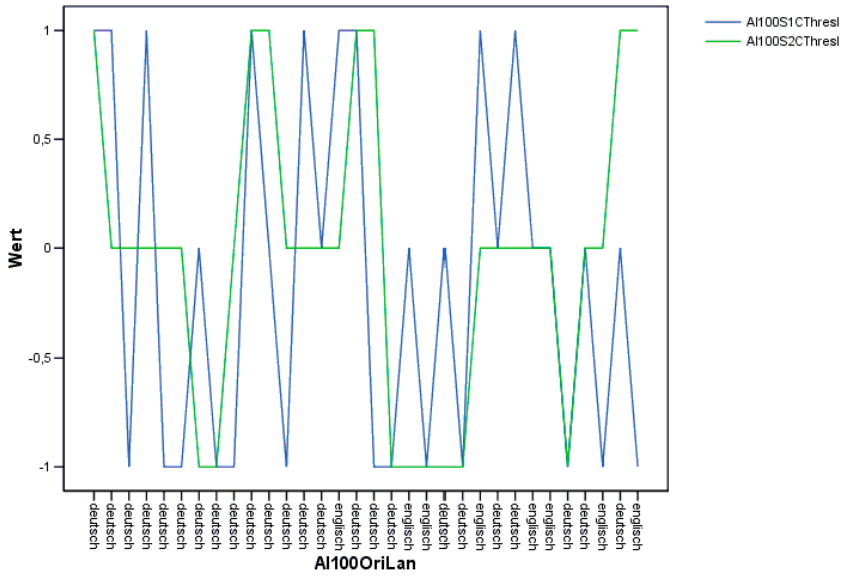
		AI100S2CRepQuo	AI100S2CThresI
AI100S2CRepQuo	Pearson Correlation	1	,874**
	Sig. (2-tailed)		,000
	Sum of Squares and Cross-products	,118	1,162
	Covariance	,004	,037
	N	32	32
AI100S2CThresI	Pearson Correlation	,874**	1
	Sig. (2-tailed)	,000	
	Sum of Squares and Cross-products	1,162	14,969
	Covariance	,037	,483
	N	32	32

**. Correlation is significant at the 0.01 level (2-tailed).

Fig. 87: Correlation between Graph of TRQ measure in AI1k corpus test sets with ThresI indicator for transgression (+1) and under run (-1) of 95% confidence intervals

In Fig. 87 the correlation between Graph of ThresI measures in AI1k corpus test sets are shown. A significant positive correlation between both measures was present. This does not mean a present dependency between source language and ThresI measure, but rather a strong relationship between the indicator of violating the lower or upper border of confidence of TRQ (ThresI) and the TRQ measure itself.

The graph of ThresI has a very different shape compared to both test sets. Fig. 88 shows both graphs together with their correlation values.

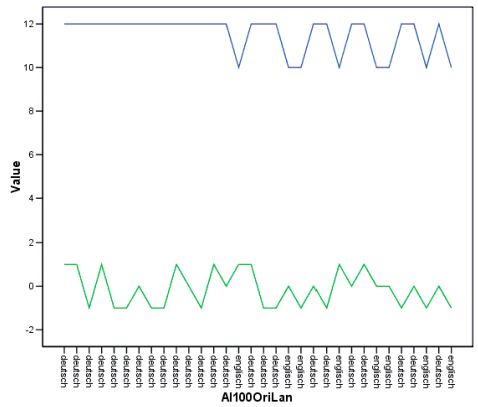


Korrelationen

		AI100S1 CThresl	AI100S2 CThresl
AI100S1CThresl	Korrelation nach Pearson	1	,327
	Signifikanz (2-seitig)		,067
	Quadratsummen und Kreuzprodukte	21,500	5,875
	Kovarianz	,694	,190
	N	32	32
AI100S2CThresl	Korrelation nach Pearson	,327	1
	Signifikanz (2-seitig)	,067	
	Quadratsummen und Kreuzprodukte	5,875	14,969
	Kovarianz	,190	,483
	N	32	32

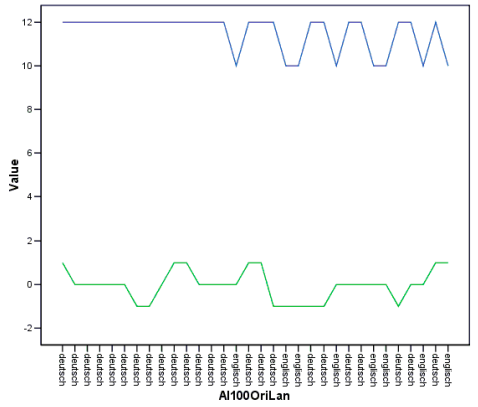
Fig. 88: Graph and correlations of indicator for transgression (+1) and under run (-1) of 95% confidence intervals of both AI1k corpus test sets

A positive but weak correlation is present. To analyse the factors that may indicate a statistical difference between yearly corpus segments with different source languages an additional indicator (LanInd) was created where value “12” equals German and value “10” equals English. Fig. 89 shows graphs and correlation analysis results between ThresI and LanInd for both test sets. There was no significant correlation found. The correlation found was very weak and only present in one test set.



Correlations

		LanInd	AI100S1 CThresI
LanInd	Pearson Correlation	1	,000
	Sig. (2-tailed)		1,000
	Sum of Squares and Cross-products	24,000	,000
	Covariance	,774	,000
	N	32	32
AI100S1CThresI	Pearson Correlation	,000	1
	Sig. (2-tailed)	1,000	
	Sum of Squares and Cross-products	,000	21,500
	Covariance	,000	,694
	N	32	32



Correlations

		LanInd	AI100S2 CThresI
LanInd	Pearson Correlation	1	,079
	Sig. (2-tailed)		,667
	Sum of Squares and Cross-products	24,000	1,500
	Covariance	,774	,048
	N	32	32
AI100S2CThresI	Pearson Correlation	,079	1
	Sig. (2-tailed)	,667	
	Sum of Squares and Cross-products	1,500	14,969
	Covariance	,048	,483
	N	32	32

Fig. 89: Correlation between Graph of Language indicator (12 eq. German; 10 eq. English) in AI1k corpus test sets with indicator ThresI for transgression (+1) and under run (-1) of 95% confidence intervals

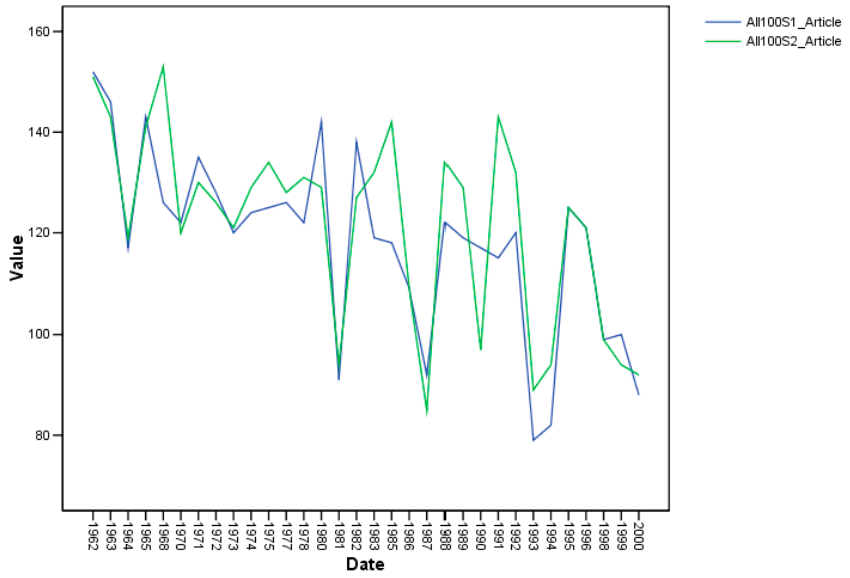
This result permits us to conclude that there is no simple aggregated measure like TRQ, LanInd or ThresI available from a statistical perspective, which indicates language of origin on corpus level.

4.4.2 Language fingerprint on corpus-level summary

- *The violation of TRQ measure confidence intervals is strongly correlated to TRQ measure value itself*
- *The violation of TRQ measure confidence intervals is not significantly correlated to both test sets*
- *There was no significant correlation found between language indicator and the indicator for violation of TRQ measure confidence intervals*

4.4.3 Language fingerprint on concept level

It is to be expected that each certain language does have a special kind of distribution of certain terms and term classes. Test sets were prepared assigning each matching term to one of the grammatical classes in “Dim_CC” taxonomy. Initially the equality of test sets was tested. Due to the use of two separate test sets it was possible to analyse the inter-test-set correlation between the same term classes. The results are shown in Fig. 90 to Fig. 95.

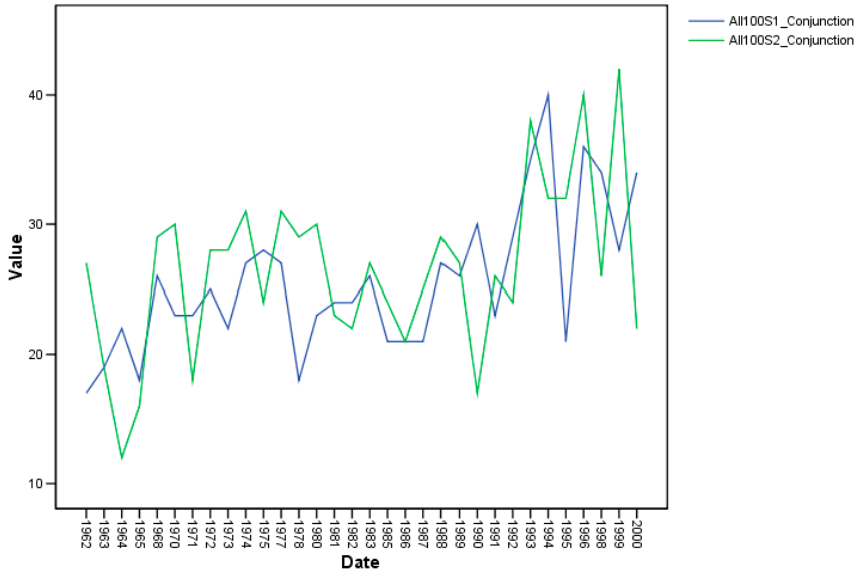


Correlations

		All100S1_ Article	All100S2_ Article
All100S1_Article	Pearson Correlation	1	,836**
	Sig. (2-tailed)		,000
	N	32	32
All100S2_Article	Pearson Correlation	,836**	1
	Sig. (2-tailed)	,000	
	N	32	32

**. Correlation is significant at the 0.01 level (2-tailed).

Fig. 90: Graphs and correlations of CountSum measure in $AI1k_{S1}$ and $AI1k_{S2}$ corpus test sets for the concept “Article” of dimension CC_Dim

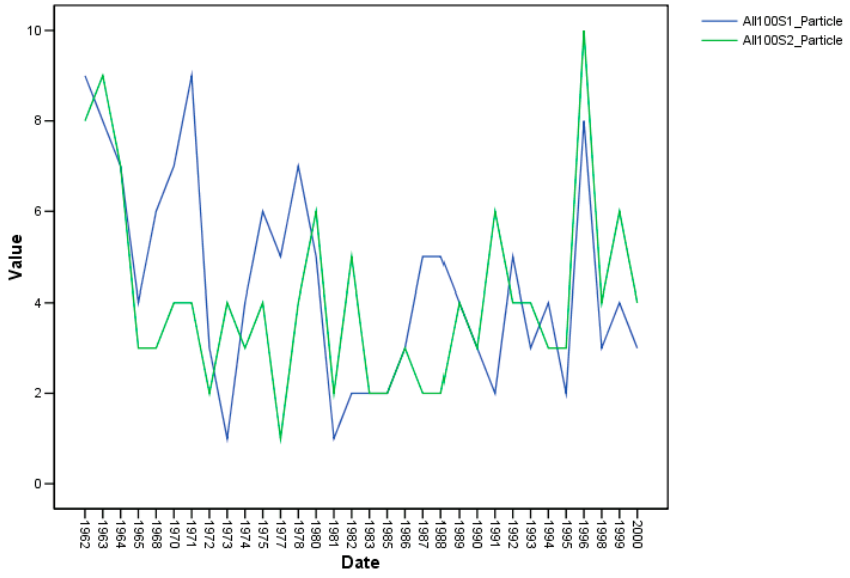


Correlations

		All100S1_ Conjunction	All100S2_ Conjunction
All100S1_Conjunction	Pearson Correlation	1	,413*
	Sig. (2-tailed)		,019
	N	32	32
All100S2_Conjunction	Pearson Correlation	,413*	1
	Sig. (2-tailed)	,019	
	N	32	32

*. Correlation is significant at the 0.05 level (2-tailed).

Fig. 91: Graphs and correlations of CountSum measure in $AI1k_{S1}$ and $AI1k_{S2}$ corpus test sets for the concept “Conjunction” of dimension CC_Dim

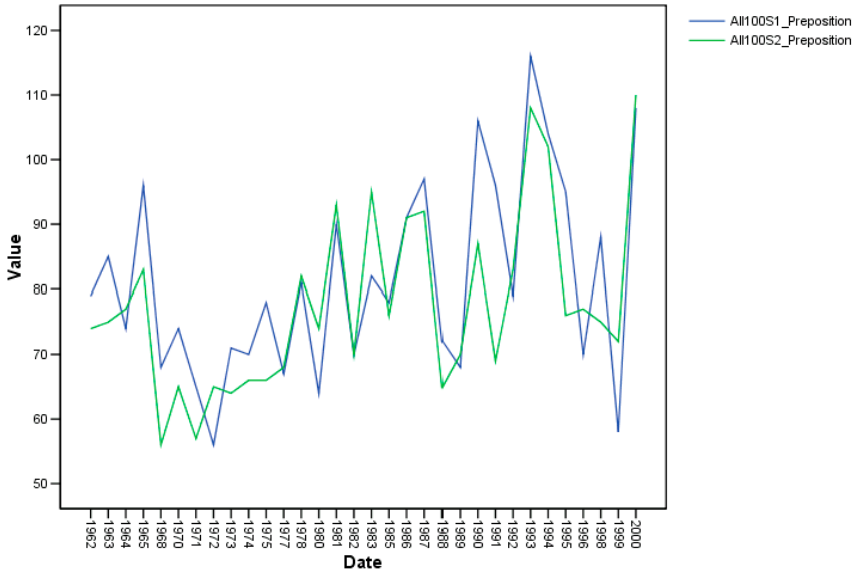


Correlations

		All100S1_ Particle	All100S2_ Particle
All100S1_Particle	Pearson Correlation	1	,528**
	Sig. (2-tailed)		,002
	N	32	32
All100S2_Particle	Pearson Correlation	,528**	1
	Sig. (2-tailed)	,002	
	N	32	32

**. Correlation is significant at the 0.01 level (2-tailed).

Fig. 92: Graphs and correlations of CountSum measure in $AI1k_{S1}$ and $AI1k_{S2}$ corpus test sets for the concept “Particle” of dimension CC_Dim

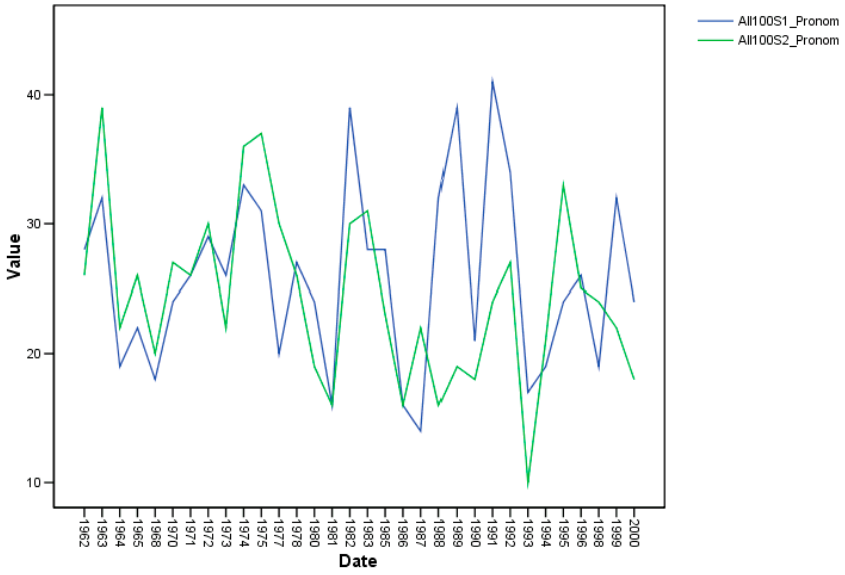


Correlations

		All100S1_ Preposition	All100S2_ Preposition
All100S1_Preposition	Pearson Correlation	1	,789**
	Sig. (2-tailed)		,000
	N	32	32
All100S2_Preposition	Pearson Correlation	,789**	1
	Sig. (2-tailed)	,000	
	N	32	32

**. Correlation is significant at the 0.01 level (2-tailed).

Fig. 93: Graphs and correlations of CountSum measure in $AI1k_{S1}$ and $AI1k_{S2}$ corpus test sets for the concept “Preposition” of dimension CC_Dim

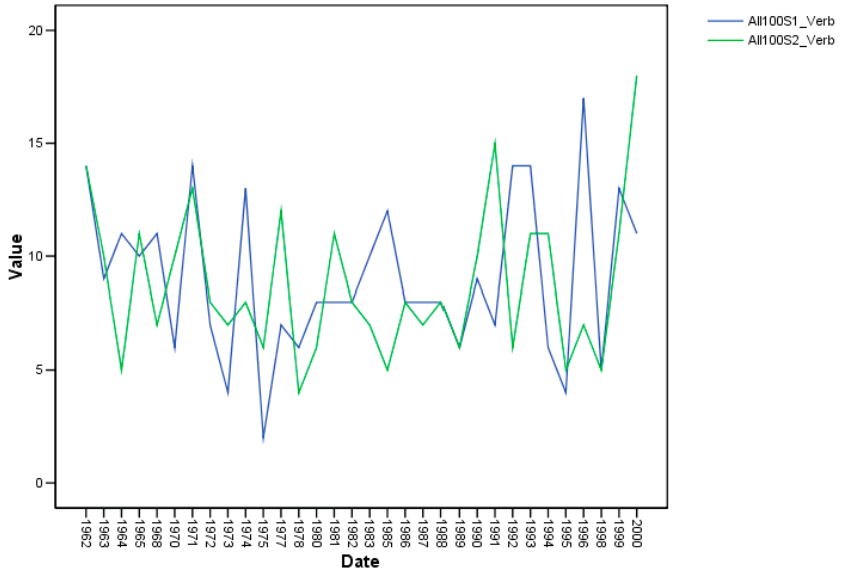


Correlations

		All100S1_ Pronom	All100S2_ Pronom
All100S1_Pronom	Pearson Correlation	1	,422*
	Sig. (2-tailed)		,016
	N	32	32
All100S2_Pronom	Pearson Correlation	,422*	1
	Sig. (2-tailed)	,016	
	N	32	32

*. Correlation is significant at the 0.05 level (2-tailed).

Fig. 94: Graphs and correlations of CountSum measure in $AI1k_{S1}$ and $AI1k_{S2}$ corpus test sets for the concept “Pronoun” of dimension CC_Dim



Correlations

		All100S1_ Verb	All100S2_ Verb
All100S1_Verb	Pearson Correlation	1	,261
	Sig. (2-tailed)		,149
	N	32	32
All100S2_Verb	Pearson Correlation	,261	1
	Sig. (2-tailed)	,149	
	N	32	32

Fig. 95: Graphs and correlations of CountSum measure in $Al1k_{S1}$ and $Al1k_{S2}$ corpus test sets for the concept “Verb” of dimension CC_Dim

Only the class “Verb” is not significantly correlated to both test sets. The dominant influence of verbs could be the reason due to the limitation of 1,000 terms per test set. It is not to be expected from this analysis that similar results will be derived for the class “Verb” in further analyses. All other classes allow conclusions in derived results due to their statistical similarity. The analysis on concept level starts with the complete set of time segments no matter whether these were translated from English or if they were of German origin.

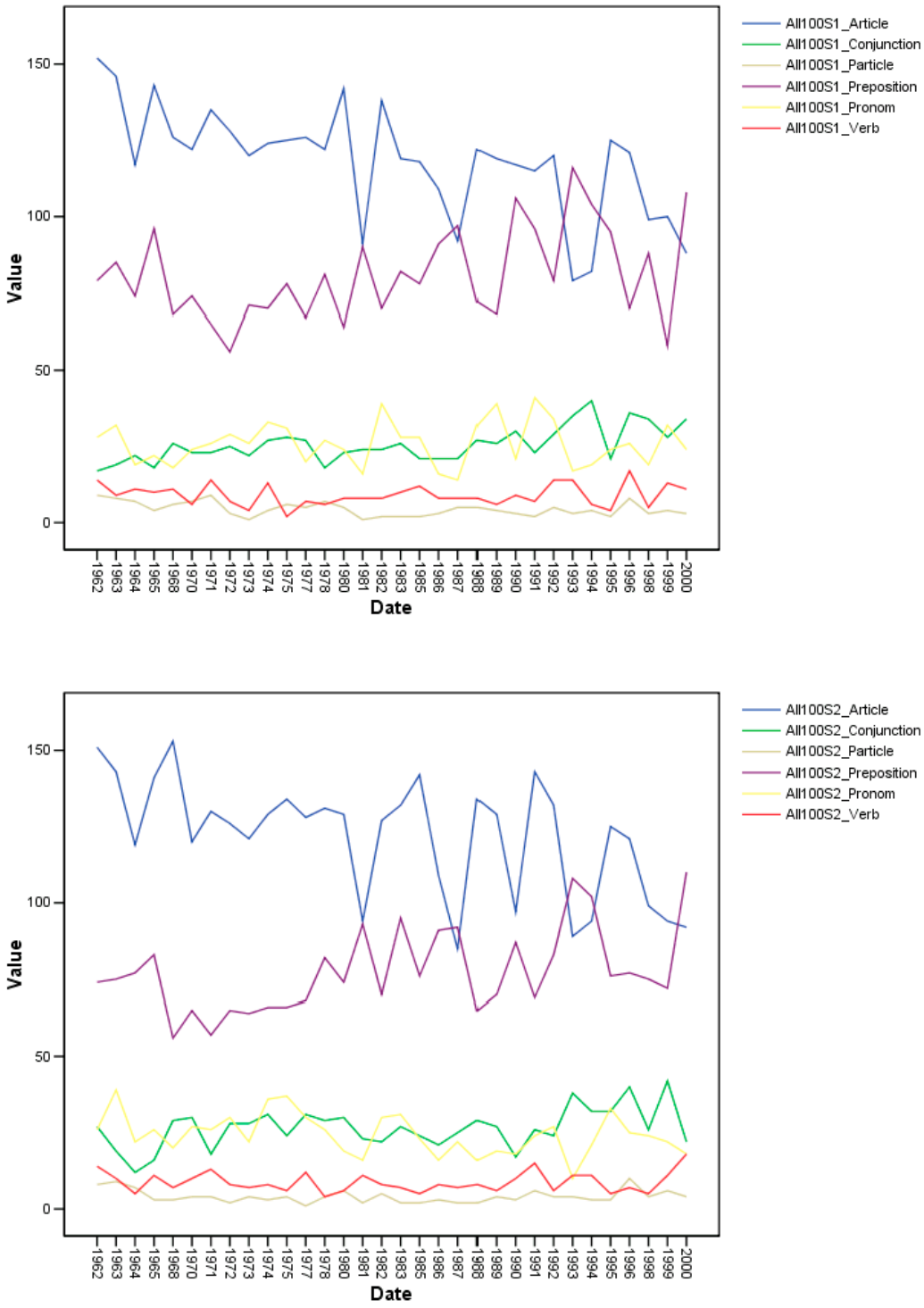


Fig. 96: Graph of CountSum measure in $Al1k_{S1}$ and $Al1k_{S2}$ corpus test sets for each single concept that constitutes the dimension CC_Dim

The number of terms that were assigned to the different grammatical classes over the observation period is shown in Fig. 96. In contrast to the observations on corpus level, term-class-level differences can be found in distributions in time. For both test sets $Al1k_{S1}$ and $Al1k_{S2}$ separate correlation analy-

ses were applied on the classes of each certain test set. Significant correlations were found in both test sets. Only the concepts “Particle” and “Verb” did not correlate with other classes and the class “Conjunction” only correlated with other concepts within one of the test sets (see Fig. 114 and Fig. 115 in Appendix).

4.4.4 Language fingerprint on concept-level summary

- *Both Allianz test sets are statistically quite similar, only the concept “Verb” showed different qualities.*
- *Significant source language (German/English) correlations were found for the concepts “Article”, “Preposition” and “Pronoun”.*

4.4.5 Analysis of statistical indicators for German corpus subsets

For a detailed analysis of the language fingerprint the AI1k test sets were separated into two subsets that contain only yearly issues of German origin, on the one hand, and of English origin, on the other. In this chapter the German subset is analysed. 24 out of 32 yearly segments were from German origin. The graphs of CountSum measure of terms assigned to “CC_Dim” concepts are shown in Fig. 97.

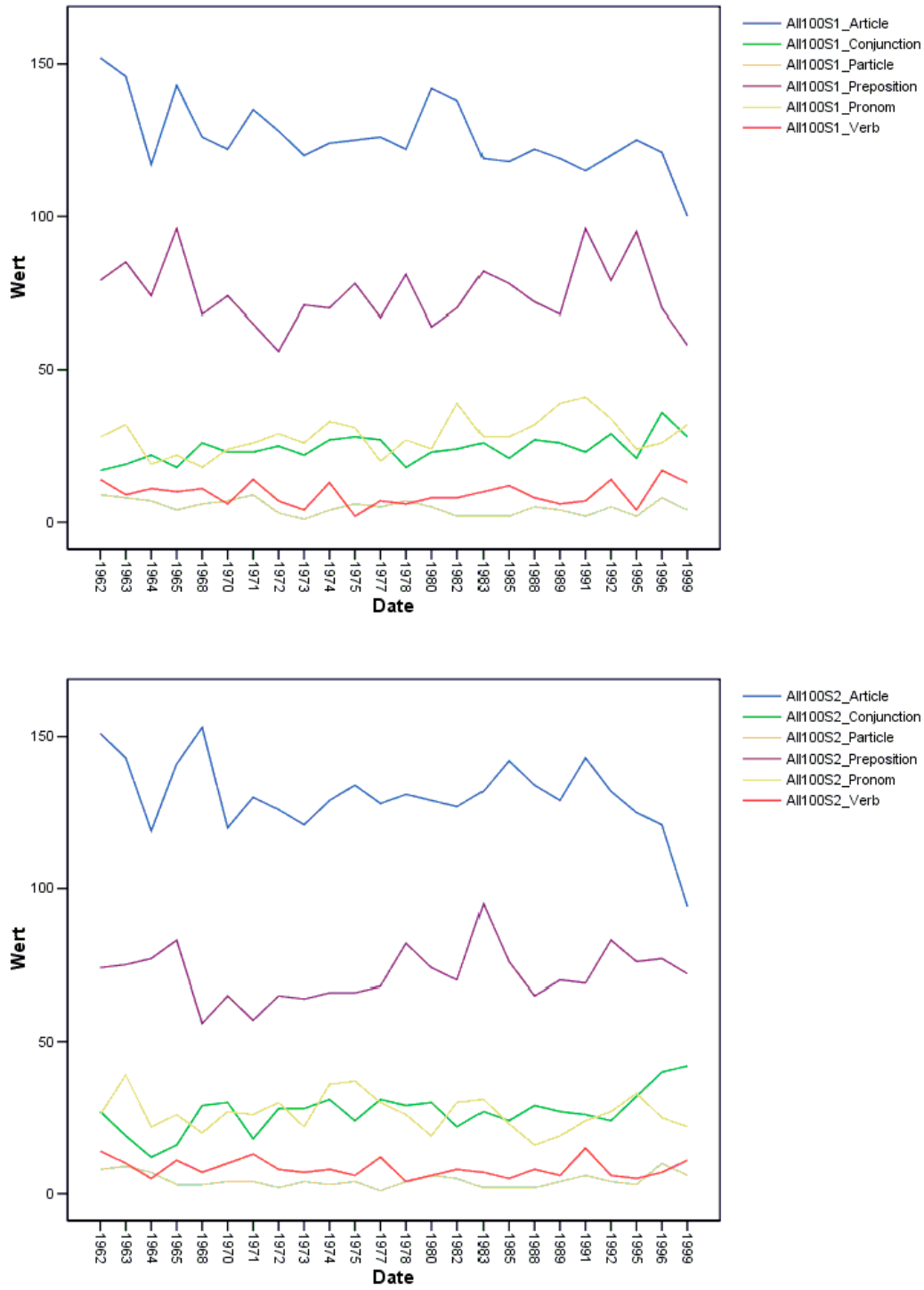
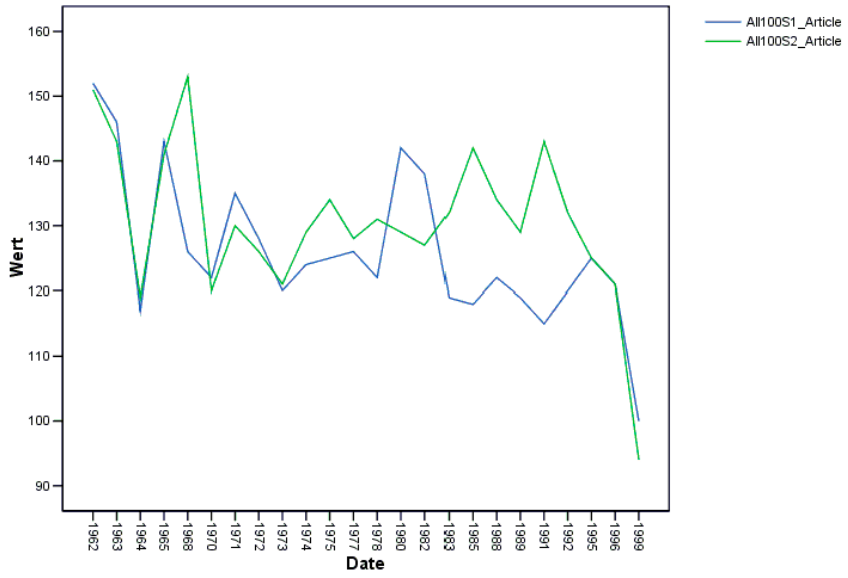


Fig. 97: Graph of CountSum measure in $AI1k_{S1}$ and $AI1k_{S2}$ German source language corpus test sets for each single concept that constitutes dimension CC_Dim

As an indicator for test set similarity between $AI1k_{S1}$ and $AI1k_{S2}$ the correlation of CountSum measure of both AI1k test sets for each single concept of "CC_Dim" is shown in Fig. 98 to Fig. 103:

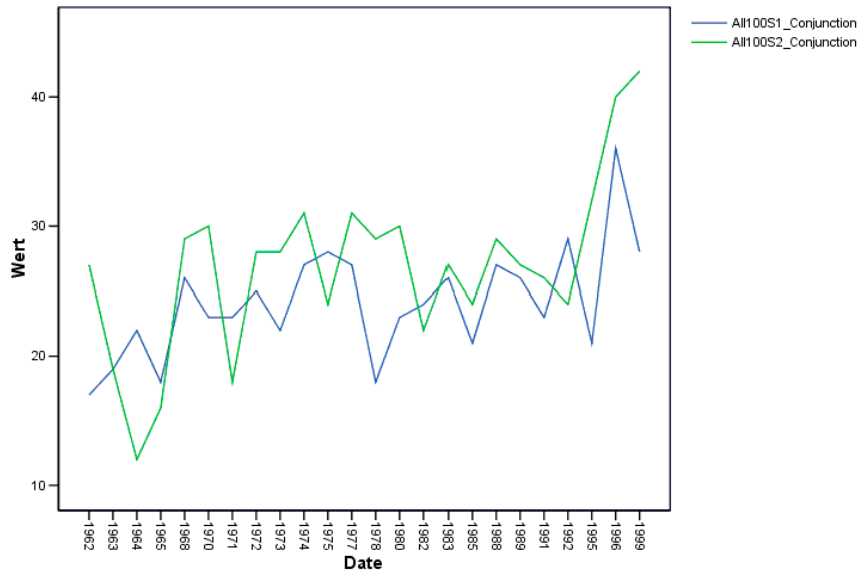


Korrelationen

		All100S1_ Article	All100S2_ Article
All100S1_Article	Korrelation nach Pearson	1	,574**
	Signifikanz (2-seitig)		,003
	N	24	24
All100S2_Article	Korrelation nach Pearson	,574**	1
	Signifikanz (2-seitig)	,003	
	N	24	24

**. Die Korrelation ist auf dem Niveau von 0,01 (2-seitig) signifikant.

Fig. 98: Graphs and correlations of CountSum measure in $Al1k_{S1}$ and $Al1k_{S2}$ German source language corpus test sets for the concept “Article” of imension CC_Dim

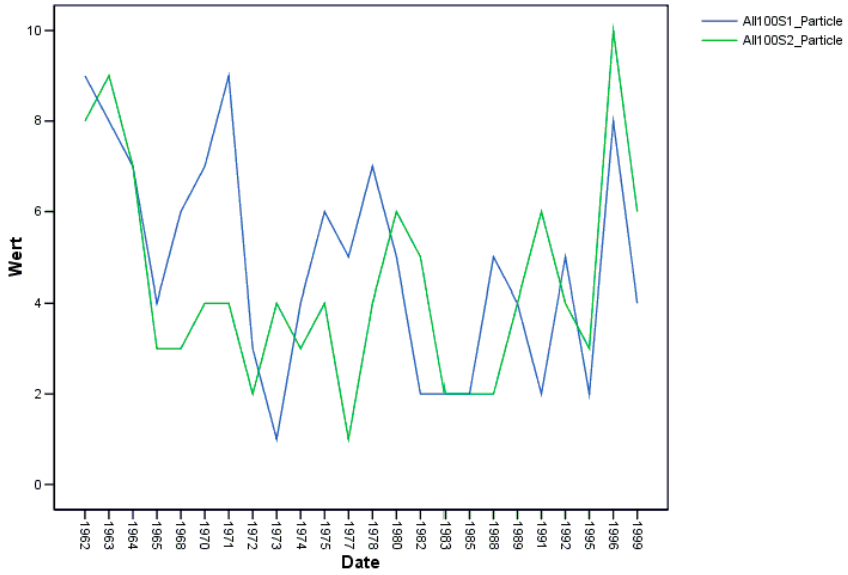


Korrelationen

		All100S1_ Conjunction	All100S2_ Conjunction
All100S1_Conjunction	Korrelation nach Pearson	1	,527**
	Signifikanz (2-seitig)		,008
	N	24	24
All100S2_Conjunction	Korrelation nach Pearson	,527**	1
	Signifikanz (2-seitig)	,008	
	N	24	24

**. Die Korrelation ist auf dem Niveau von 0,01 (2-seitig) signifikant.

Fig. 99: Graphs and correlations of CountSum measure in $AI1k_{S1}$ and $AI1k_{S2}$ German source language corpus test sets for the concept “Conjunction” of dimension CC_Dim

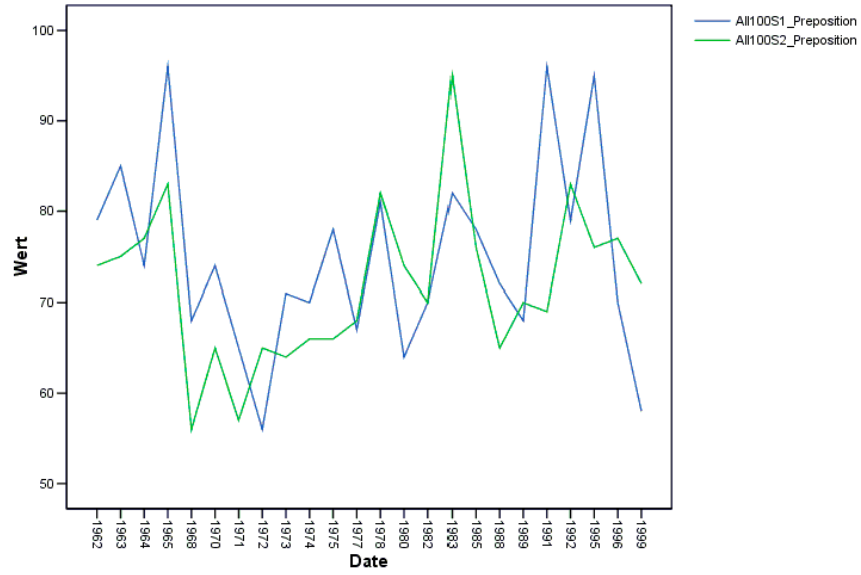


Korrelationen

		All100S1_ Particle	All100S2_ Particle
All100S1_Particle	Korrelation nach Pearson	1	,509*
	Signifikanz (2-seitig)		,011
	N	24	24
All100S2_Particle	Korrelation nach Pearson	,509*	1
	Signifikanz (2-seitig)	,011	
	N	24	24

*. Die Korrelation ist auf dem Niveau von 0,05 (2-seitig) signifikant.

Fig. 100: Graphs and correlations of CountSum measure in $Al1k_{S1}$ and $Al1k_{S2}$ German source language corpus test sets for the concept “Particle” of dimension CC_Dim

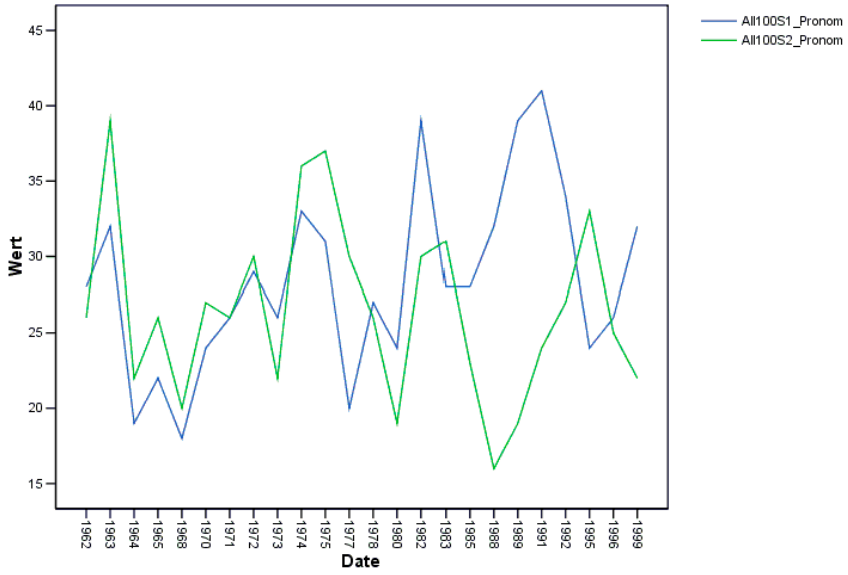


Korrelationen

		All100S1_ Preposition	All100S2_ Preposition
All100S1_Preposition	Korrelation nach Pearson	1	,478*
	Signifikanz (2-seitig)		,018
	N	24	24
All100S2_Preposition	Korrelation nach Pearson	,478*	1
	Signifikanz (2-seitig)	,018	
	N	24	24

*. Die Korrelation ist auf dem Niveau von 0,05 (2-seitig) signifikant.

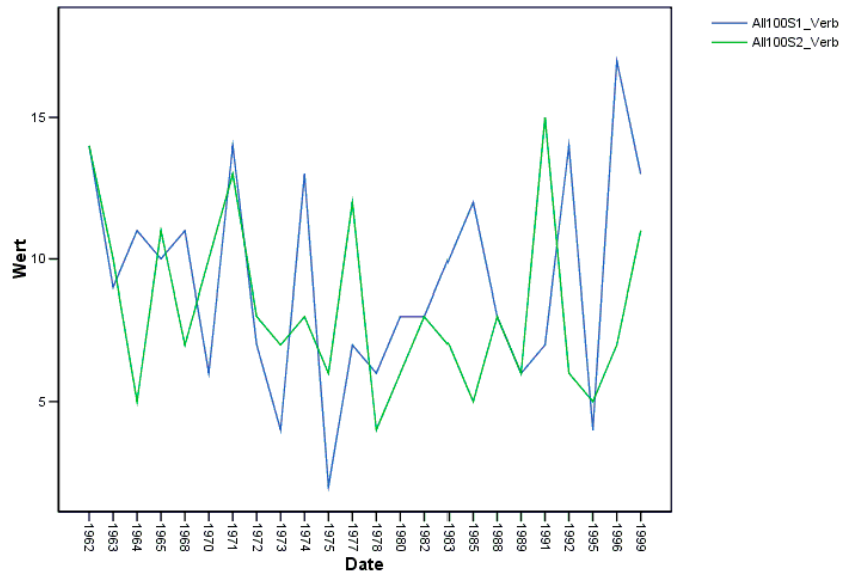
Fig. 101: Graphs and correlations of CountSum measure in $Al1k_{S1}$ and $Al1k_{S2}$ German source language corpus test sets for the concept “Preposition” of dimension CC_Dim



Korrelationen

		All100S1_ Pronom	All100S2_ Pronom
All100S1_Pronom	Korrelation nach Pearson	1	,114
	Signifikanz (2-seitig)		,597
	N	24	24
All100S2_Pronom	Korrelation nach Pearson	,114	1
	Signifikanz (2-seitig)	,597	
	N	24	24

Fig. 102: Graphs and correlations of CountSum measure in $Al1k_{S1}$ and $Al1k_{S2}$ German source language corpus test sets for the the concept “Pronoun” of dimension CC_Dim



Korrelationen

		All100S1_ Verb	All100S2_ Verb
All100S1_Verb	Korrelation nach Pearson	1	,221
	Signifikanz (2-seitig)		,300
	N	24	24
All100S2_Verb	Korrelation nach Pearson	,221	1
	Signifikanz (2-seitig)	,300	
	N	24	24

Fig. 103: Graphs and correlations of CountSum measure in $Al1k_{S1}$ and $Al1k_{S2}$ German source language corpus test sets for the concept “Verb” of dimension CC_Dim

The correlations between the concepts were generally positive for all concepts. Significance was not found for the concepts “Pronoun” and “Verb”. For these concepts the test sets do not reflect the same occurrence. Table 66 (see Appendix) mentions the correlations between CountSum measures in Al1k German source language corpus test sets for each single concept of “CC_Dim”.

In both test sets only one significant correlation was found: A negative correlation between “Article” and “Conjunction”. This result indicates a substitution use of both grammatical concepts in Al1k test sets.

It is to be expected that the frequency of use of certain concept is typical for languages. Based on the German subset ranked lists of concepts for all test sets with different intensity of pre-processing were calculated (see Fig. 104):

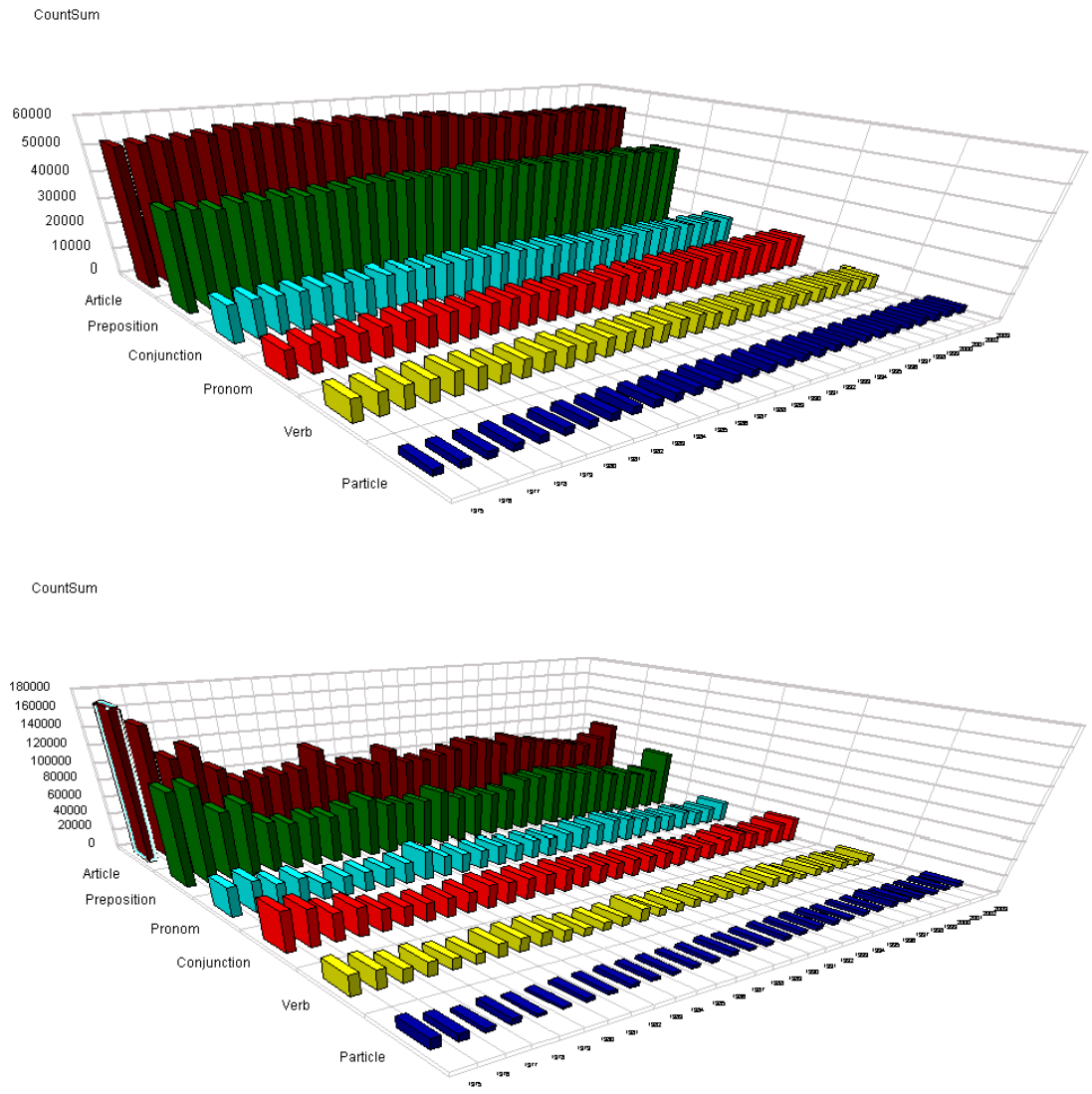


Fig. 104: Examples of ranked lists of CountSum measure from concepts of CC_Dim as subsets from the CW_{5k} and CW_{5kb} test sets

For all type “n” CW test sets the ranked order of use was found as: *Article, Preposition, Conjunction, Pronoun, Verb, Particle*.

In type “b” CW test sets the order was found as:
Article, Preposition, Pronoun, Conjunction, Verb, Particle.

The Al1k test sets did not deliver equal results, “Pronoun” and “Conjunction” changed their ranks between both test sets, so the results derived from Al1k test sets were not typical, but between the other analysed type “n” and type “b” test sets. Probably this result occurred due to the limited quantity of terms of a number of 1,000 per yearly segment.

4.4.6 Analysis of statistical indicators for German corpus subsets summary

- *For Al1k test set a significant negative correlation was found between the the concepts “Article” and “Conjunction”, which indicates a substitution use of both grammatical concepts within the test sets.*
- *For all type “n” CW test sets the ranked order of used concepts was found as: Article, Preposition, Conjunction, Pronoun, Verb, Particle.*
- *For all type “b” and “bn” CW test sets the order of used concepts was found as: Article, Preposition, Pronoun, Conjunction, Verb, Particle.*
- *For Al1k test set the calculation of rank lists led to no equal results for both test sets with a realization between the results from CW test sets type “n” and “b/n”.*

4.4.7 Analysis of statistical indicators for Al1k English corpus subsets

In this chapter the English subset is analysed. 8 out of 32 yearly segments were of English origin. This is a small base for statistically relevant results, but it can provide a rough orientation. The graphs of CountSum measure of terms assigned to the concepts of “CC_Dim” are shown in Fig. 105.

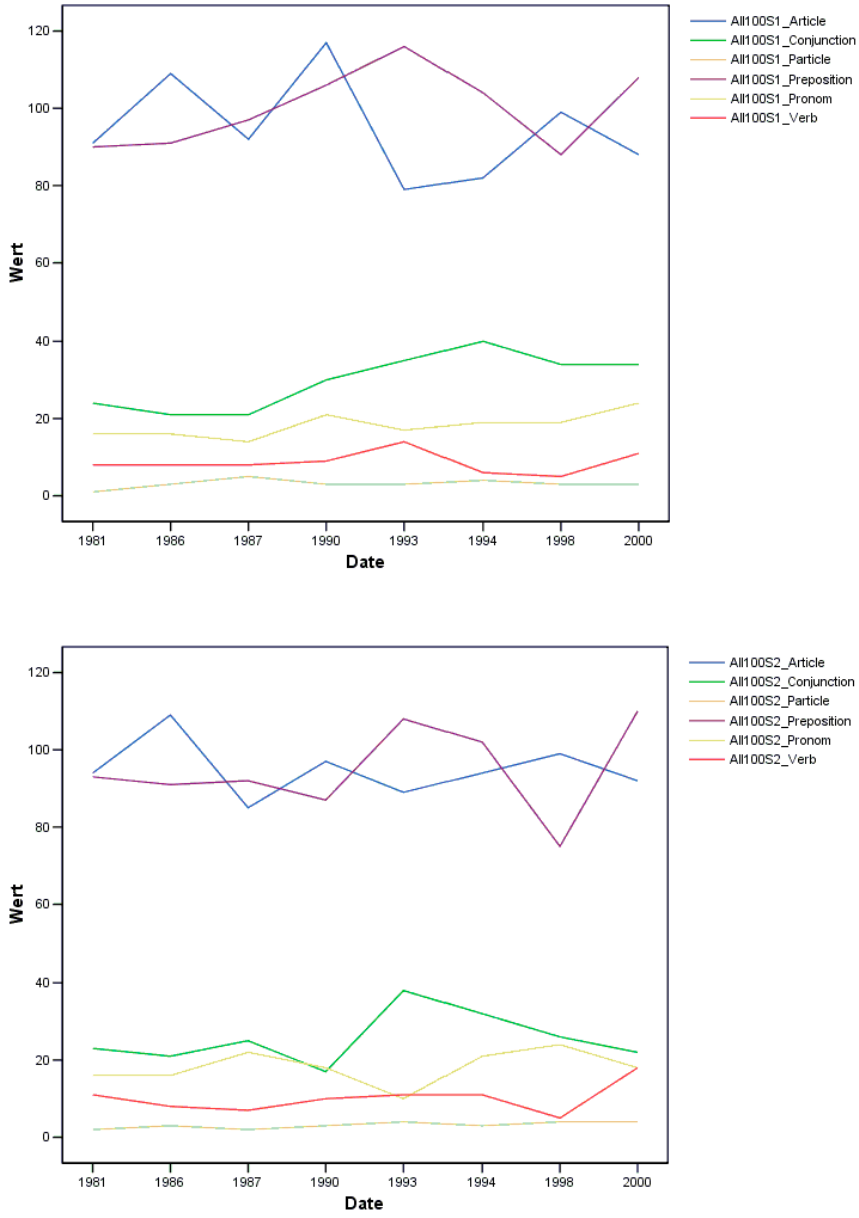


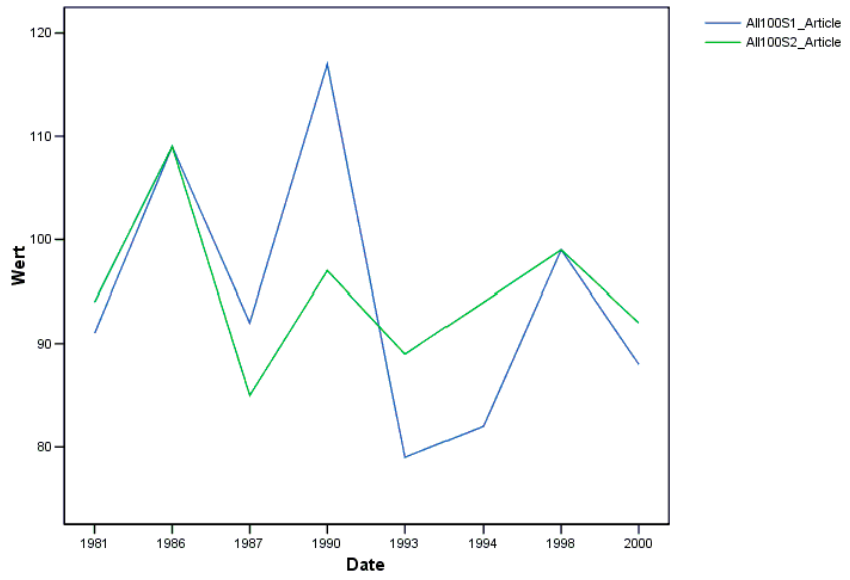
Fig. 105: Graph of CountSum measure in $Al1k_{S1}$ and $Al1k_{S2}$ English source language corpus test sets for each single concept that constitutes dimension CC_Dim

Table 67 mentions the correlations between LanInd and CountSum measure in Al1k English source language corpus test set for each single concept of

“CC_Dim”. Contrary to the German subsets in both test sets no significant correlations were found.

The absence of a significant correlation indicates a difference in use of the grammatical concepts between the German and the English corpus subsets.

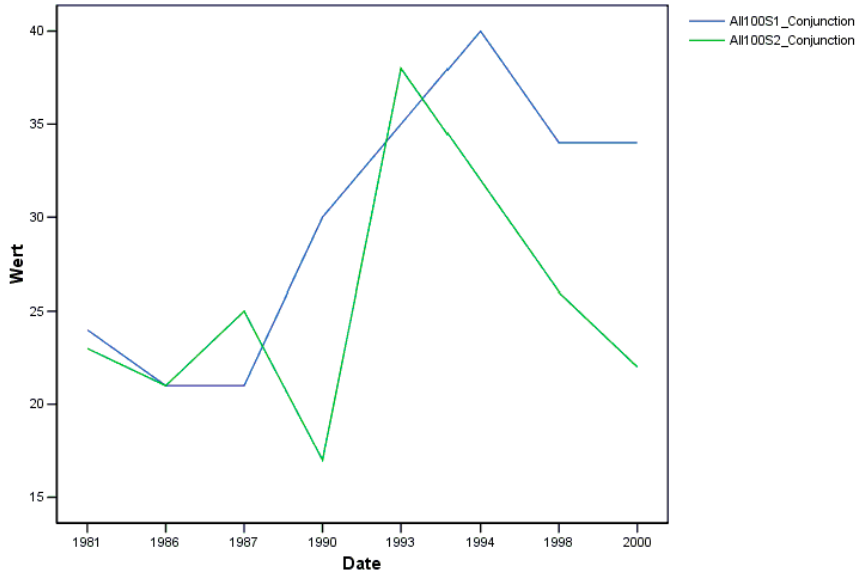
As an indicator for test set similarity the correlation of CountSum measure of both Al1k test sets for each single concept of “CC_Dim” is shown in Fig. 106 to Fig. 111:



Korrelationen

		All100S1_ Article	All100S2_ Article
All100S1_Article	Korrelation nach Pearson	1	,638
	Signifikanz (2-seitig)		,089
	N	8	8
All100S2_Article	Korrelation nach Pearson	,638	1
	Signifikanz (2-seitig)	,089	
	N	8	8

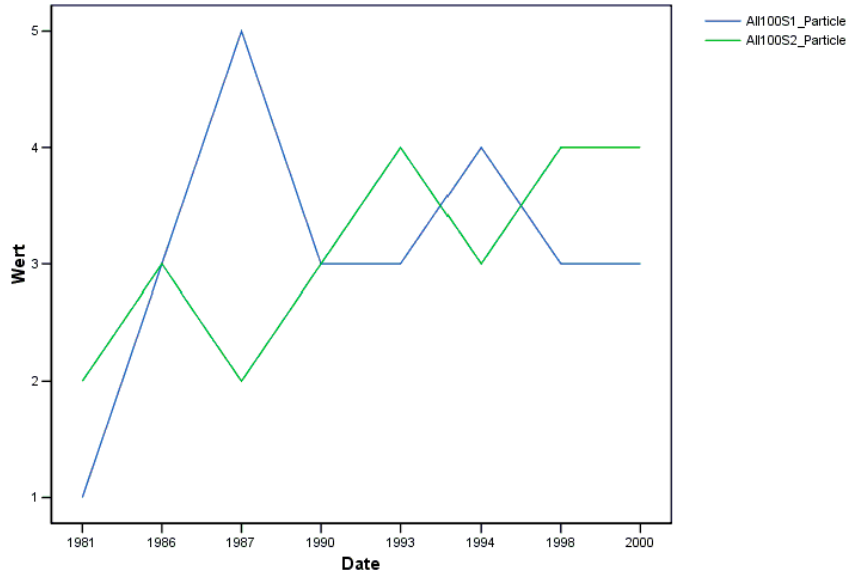
Fig. 106: Graphs and correlations of CountSum measure in $Al1k_{S1}$ and $Al1k_{S2}$ English source language corpus test sets for the concept “Article” of dimension CC_Dim



Korrelationen

		All100S1_ Conjunction	All100S2_ Conjunction
All100S1_Conjunction	Korrelation nach Pearson	1	,529
	Signifikanz (2-seitig)		,178
	N	8	8
All100S2_Conjunction	Korrelation nach Pearson	,529	1
	Signifikanz (2-seitig)	,178	
	N	8	8

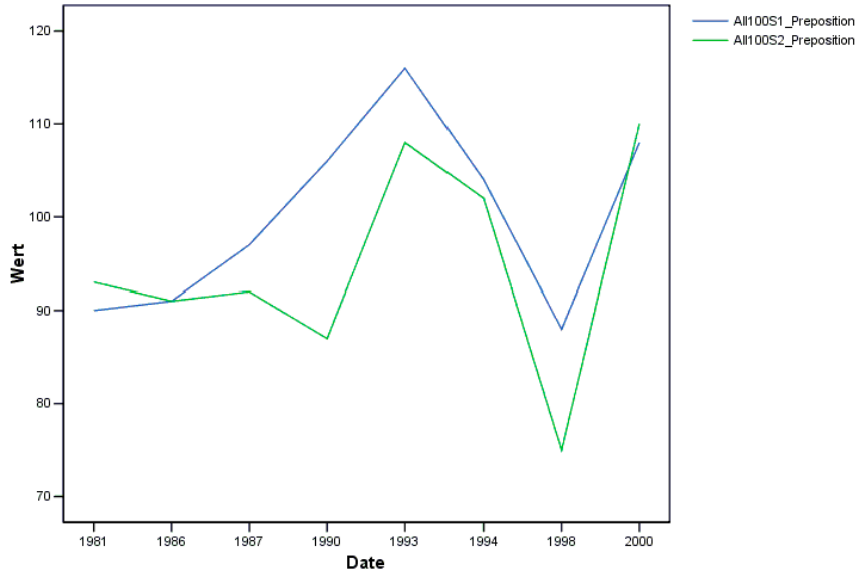
Fig. 107: Graphs and correlations of CountSum measure in $AI1k_{S1}$ and $AI1k_{S2}$ English source language corpus test sets for the concept “Conjunction” of dimension CC_Dim



Korrelationen

		All100S1_ Particle	All100S2_ Particle
All100S1_Particle	Korrelation nach Pearson	1	-,019
	Signifikanz (2-seitig)		,964
	N	8	8
All100S2_Particle	Korrelation nach Pearson	-,019	1
	Signifikanz (2-seitig)	,964	
	N	8	8

Fig. 108: Graphs and correlations of CountSum measure in $Al1k_{S1}$ and $Al1k_{S2}$ English source language corpus test sets for the concept “Particle” of dimension CC_Dim

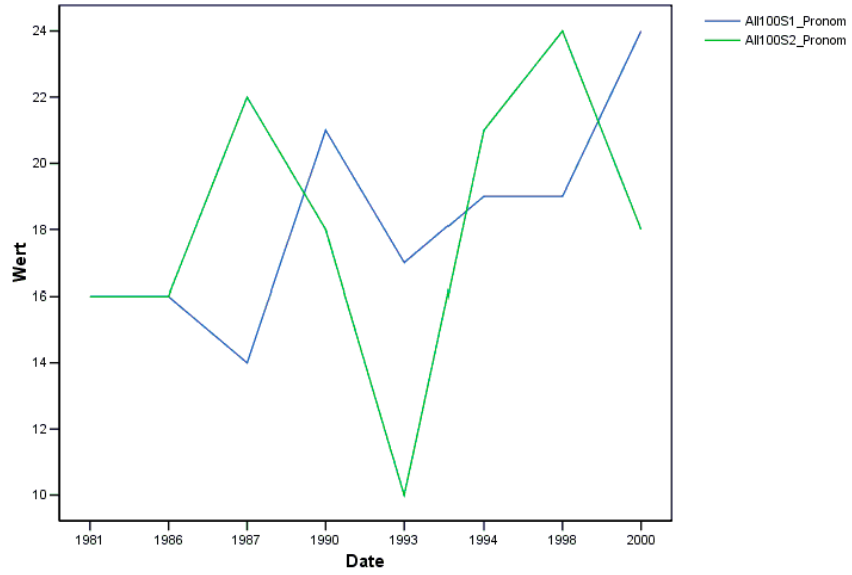


Korrelationen

		All100S1_ Preposition	All100S2_ Preposition
All100S1_Preposition	Korrelation nach Pearson	1	,754*
	Signifikanz (2-seitig)		,031
	N	8	8
All100S2_Preposition	Korrelation nach Pearson	,754*	1
	Signifikanz (2-seitig)	,031	
	N	8	8

*. Die Korrelation ist auf dem Niveau von 0,05 (2-seitig) signifikant.

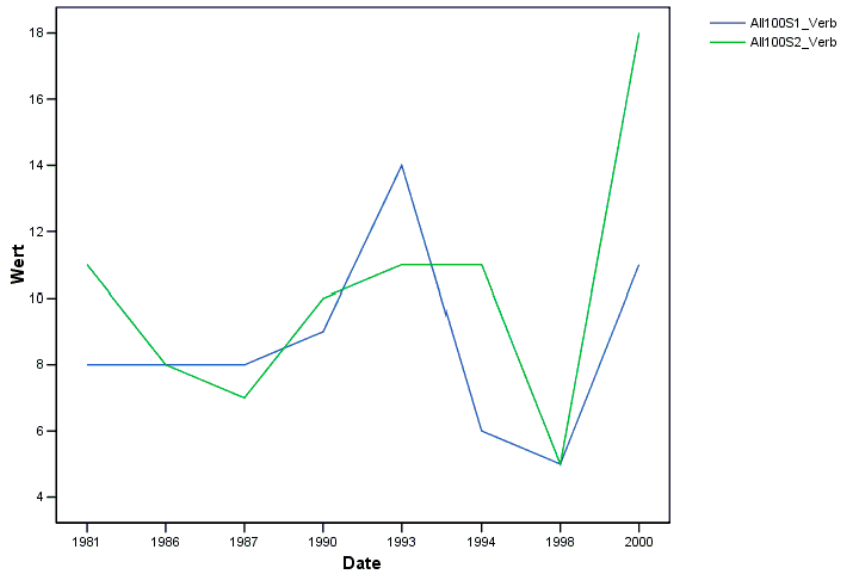
Fig. 109: Graphs and correlations of CountSum measure in $Al1k_{S1}$ and $Al1k_{S2}$ English source language corpus test sets for the concept “Preposition” of dimension CC_Dim



Korrelationen

		All100S1_ Pronom	All100S2_ Pronom
All100S1_Pronom	Korrelation nach Pearson	1	,090
	Signifikanz (2-seitig)		,833
	N	8	8
All100S2_Pronom	Korrelation nach Pearson	,090	1
	Signifikanz (2-seitig)	,833	
	N	8	8

Fig. 110: Graphs and correlations of CountSum measure in $Al1k_{S1}$ and $Al1k_{S2}$ English source language corpus test sets for the concept “Pronoun” of dimension CC_Dim



Korrelationen

		All100S1_ Verb	All100S2_ Verb
All100S1_Verb	Korrelation nach Pearson	1	,554
	Signifikanz (2-seitig)		,155
	N	8	8
All100S2_Verb	Korrelation nach Pearson	,554	1
	Signifikanz (2-seitig)	,155	
	N	8	8

Fig. 111: Graphs and correlations of CountSum measure in $AI1k_{S1}$ and $AI1k_{S2}$ English source language corpus test sets for the concept “Verb” of dimension CC_Dim

The correlations between the concepts were generally positive for all concepts, but not for “Particle”. Significance was only found for the concept “Preposition”. For all other concepts the test sets do not reflect the same occurrence, but a high volatility. Both test sets are, in terms of a statistical analysis, not very similar.

For the English subset rank lists[35] were calculated for both test sets (see Fig. 112):

[35] based on mean CountSum values considering all periods

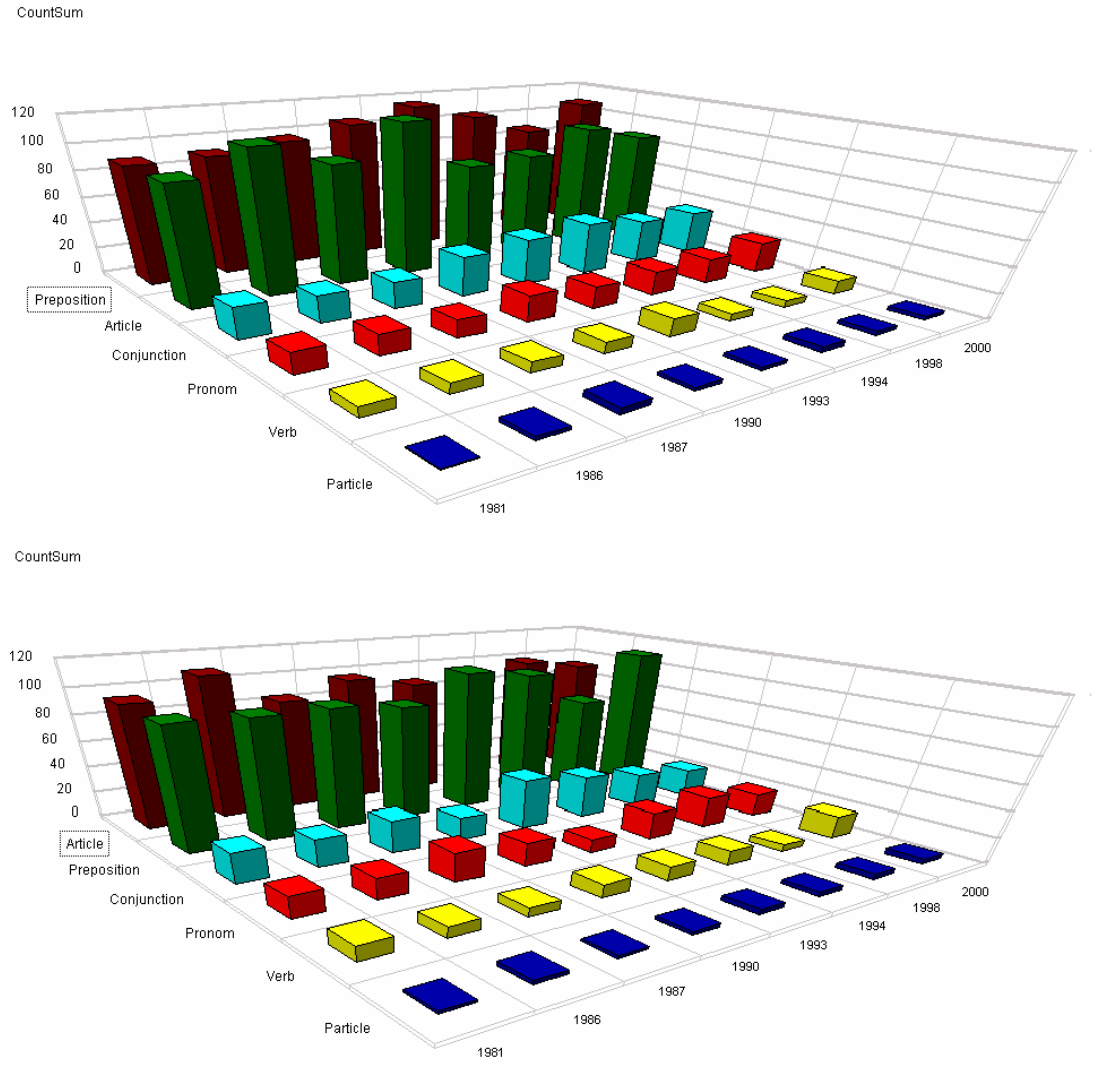


Fig. 112: Ranked list of CountSum measure from concepts of CC_Dim as subsets from $Al1k_{S1}$ $Al1k_{S2}$ test sets

The concepts “Preposition” and “Article” occur with almost the same frequency. There was only one difference between the test sets $Al1k_{S1}$ and $Al1k_{S2}$ that led to a switch in the order of concepts.

For English the results in both type “n” Al1k test sets are as follows:

$Al1k_{S1}$: *Preposition, Article, Conjunction, Pronoun, Verb, Particle*.

$Al1k_{S2}$: *Article, Preposition, Conjunction, Pronoun, Verb, Particle*.

The Al1k test sets did not deliver equal results: “Pronoun” and “Conjunction” changed their ranks between both test sets, so the results derived from Al1k test sets were not typical for other Al1k test sets, but were typical for the other analysed type “n” and type “b” test sets.

4.4.8 Analysis of statistical indicators for English corpus subsets summary

- *The statistical similarity (based on a correlation analysis of CountSum measure of similar concepts) was found not to be as significant as between both Al1k test sets.*
- *For Al1k test set a significant correlation between different concepts was found.*
- *For Al1k test set the calculation of rank lists led to no identical results for both test sets with a realization between the results from CW test sets type "n" and "b/n". The results were: Al1k$_{S1}$: Preposition, Article, Conjunction, Pronoun, Verb, Particle. Al1k$_{S2}$: Article, Preposition, Conjunction, Pronoun, Verb, Particle.*

4.5 Evaluating the impact of corpus length

Corpus length dependency is a determining factor on knowledge extraction that cannot be analysed separately, but only in unity with other factors. In this chapter the experiences and statistical results of analysis that resulted from corpus length effects will be summarized.

The following perspectives are potentially interesting for the analysis:

1. Which statistical qualities of a corpus change when the corpus size is adjusted?
2. How is the quality of extracted knowledge changed when the size of a corpus is shortened?
3. Is there a “minimal” corpus size for the application of the TRQ measure threshold approach?

For questions 1.) and 2.) the test-set combination CW_{5k}/CW_{1k} is applicable, due to the similarity of source data. Only the number of token is shortened in the generation process of CW_{1k} out of CW_{5k}. The test sets $Al1k_{S1}/Al1k_{S2}$ represent very limited corpora with a number of 1,000 token per yearly segment and which match.

4.5.1 Effects on statistical qualities and their measures

The corpus size alone mostly affected direct scaling factors such as *V (N)* and also influenced basic statistical measures, e.g., standard deviation. Ceteris paribus, a smaller corpus size, led to lower standard deviation values.

Comparing different sizes of type “n” corpora (CW_{5k} and CW_{1k}, see Table 17) the results derived from large corpus sizes led to more significant results.

Dramatically more than the adjustment of the test-set size was generated from one source; the statistical qualities were influenced by the intensity of pre-processing of the source text collection. This aspect will be focused upon within the next chapter.

4.5.2 Effects on quality of extracted knowledge

Potentially, larger text may contain more information and reflected knowledge about a domain. The results of the semantic analysis documented for CW_{5k} and CW_{1k} with different corpus length are documented in Chapters 4.3.1.5.1 and 4.3.1.5.2. The extracted results were similar, but were more precise the more terms the corpus contained.

4.5.3 The “minimal” corpus size for the TRQ measure threshold approach

The TRQ measure threshold approach worked well with corpus sizes of 100,000 and 500,000 terms per yearly segment. Significant differences in extracted results were found with the very limited corpus test sets $Al1k_{S1}$ and $Al1k_{S2}$ with their size of 1,000 terms per yearly segment. Overall the extracted knowledge represented was reliable in the “highlights” of the progress of the company Allianz, but only very few were precise in rare events that partly got lost.

4.6 Evaluating the impact of knowledge domain and document source

CW and AI1k represent different corpora from two perspectives: Knowledge domain and technical kind of source. During the pre-processing of the data within the TMF process the source-independent ASCII format was used for further processing. The TRQ measure threshold approach was applied after this conversion took place. Due to this standard format all later steps can be applied equally. A source dependency in processing or knowledge extraction was not found. Independent from domain specifics progress paths for aggregated concepts and single terms were found. Neither an impact of knowledge domain nor an impact of document source was present. Differences in quality of results were identified to be result of different corpus length of the test sets and the intensity of pre-processing which was applied on them.

5 Domain knowledge interaction

Within the previous analysis a simple DM algorithm was applied. In this chapter a prototypical realization of a multidimensional representation of the extracted knowledge will be introduced. This part of the TMF allows visualizing, but also interaction with the knowledge extracted from the domain-related corpus interactively. The prototype was implemented using an OLAP tool available as a commercial product or free software. With the use of the OLAP technique the “Trend Landscape” metaphor became reality.

The basis for this exemplarily introduction of the principle methods of TMF knowledge interaction was the results extracted from the CW_{5k} corpus. To impartially see which concepts dominated the time periods, the application of the “No suffix” dimensions (see Table 9) in combination with threshold measure “ThresU”, will now be analysed.

5.1 Navigating the constant concepts of CW_{5k}

The interaction starts with the constant concepts ($\in C_C$) from CW_{5k}. The navigation follows the (dis-)aggregation path: Start node → Dim → Vendor → Term Level (see Chapter 4.2.1). The succession of the measure CountThresU along all time segments can be seen in Fig. 116 (see Appendix).

On every disaggregation level within the taxonomy the application of the threshold CountThresU leads to an aggregated representation of the levels below the actual shown level. The aggregation function is “Sum” for the measure CountThresU within this example. A drill-down along the disaggregation path of the Dim taxonomy leads to the detailed concepts shown in Fig. 117 (see Appendix).

In Fig. 117 the significant leading concepts within all time segments are made visible. The concepts appear ranked in reverse order (rank one is placed in the background on the left side). On aggregated level the concept “OS” was significantly more frequent within the CW_{5k} corpus in only a single year: 1975. Other concepts appear with different shapes of volatility.

The disaggregation within the concept “Vendor” can be seen in Fig. 118 (see Appendix).

On the left side a rank list was dynamically built out of the individual occurrence of each vendor over the periods. Vendor “IBM” dominates all time periods. It must be kept in mind that the corpus segment of CW_{5k} that is visualised here only contains concepts, which were present within all time segments. Only the application of the threshold measure CountThresU leads to a filtering of concepts that did not significantly dominate certain time periods. If an alternative measure without a threshold capability is used (e.g., Count), the visualisation becomes an imprecise set of data with less summarizing of information and less focusing on important facts.

When focusing on single year slices (e.g., 1975) the visual difference of the simple visualisation of all counts of concepts and the concept filtering based on TRQ thresholds can be easily seen when comparing Fig. 119 without and Fig. 120 (see Appendix for both) with threshold filtering. The application of the TRQ threshold allows an intuitive exploration of the concept “Vendor” along the time dimension (see Fig. 120 to Fig. 122, refer to Appendix). The rank list and the number of shown concepts are dynamically adopted based on the TRQ threshold of each single concept.

With this exploration method it is easy to recognize that the concept “Vendor” within the reflected domain IT in this corpus shows a characteristic progress path: In 1975, only eight vendors were significant out of C_C. A concentration of ten significant vendors in 1988 to only five in the last observation period has taken place.

5.2 Navigating the volatile concepts of CW_{5k}

The start node of the volatile concepts ($\in C_V$) from CW_{5k} is shown in Fig. 123 (see Appendix).

In Fig. 124 (see Appendix) the 2nd level of “Dim” taxonomy for each yearly segment with TRQ threshold filtering can be seen. This drill-down gives an overview of the order and progress of significant concepts over all time peri-

ods. For example the concepts “Geography” and “Currency” were extremely volatile over time and both had a peak in the 1990s.

Fig. 125 to Fig. 127 (see Appendix for both) provides yearly slices with ranked lists of aggregated concepts of the 2nd level within the “Dim” taxonomy for “1975”, “1988” and “2003”.

The 3rd level of drill-down is reached when navigating down the “Vendor” node. Fig. 128 to Fig. 130 (see Appendix for both) shows the significant concepts of the “Vendor” node within selected time slices. For an optimised screen output under the node “others” all concepts which did not belong to the 80% most occurring concepts were subsumed.

The progress of the concept “Vendor” (the opposite of the concepts assigned to the C_C segment) shows continuous diversification. More vendors became significantly important within the domain of IT, reflected within the CW_{5k} corpus.

6 Conclusion and perspective

The focus of this final chapter is to subsume the “lessons learned” and to give an outlook on further research that may extend the ideas of this dissertation.

In Chapter 4 various perspectives on corpora with different pre-processing intensities were applied. But how can the results are subsumed in a form so this may be a guide for a researcher at the beginning of his or her pre-processing step within the introduced TMF process? The evaluation in Chapter 4 was divided into statistical (quantitative) and semantic (qualitative) analyses. In combination this approach permits conclusions on the effect of certain statistic measure characteristics of corpora on the quality of the extracted knowledge. Dramatically differences in the knowledge extracted from corpora with optimal pre-processing intensity and corpus size were recognized (corpus type “n”) and the opposite, corpus type “b” with low pre-processing intensity. In the following, indicators will be summarized that allow one to distinguish between the theoretical corpus types introduced in Chapter 4.3.

On the basis of quite optimal pre-processed corpus test sets of type n it was possible to find reproducible results in several kinds of perspectives. By segmenting a corpus C into two corpus segments C_C and C_V the opportunity is given to correlate TRQ graphs of both segments. This analysis showed negative correlations in every test set. The significance rose with corpus size. From a DM perspective this indicates a clustering capability of the segmenting approach according to the statistical quality of TRQ time series. The knowledge that was extracted using taxonomy assignment of terms and TRQ thresholds led to best results for those test sets that had the most diversely clustered (most negative correlation for TRQ) corpus segments down to term level. For both Al1k test sets that did not have very well clustered C_C and C_V segments the extracted knowledge on term level was not comprehensible. Concluding from the results of this analysis, the significant negative correlation of TRQ values of C_C and C_V segments of one corpus test set allowed forecasting a reliable result in extracted knowledge from these test sets.

Fig. 113 gives a general overview of the discovered statistical dependencies between different levels of pre-processing and the influence of statistical qualities of the corpus segments C, C_V and C_C.

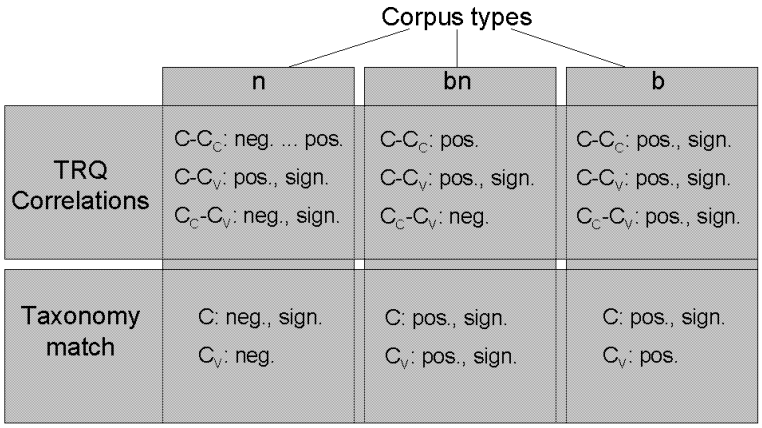


Fig. 113: Overview of different intensities of pre-processing and their statistical indicators

In the first row of Fig. 113 the correlations of TRQ of each single corpus segment are shown for the different corpus types. In the second row the matching quality between the number of different terms and the number of terms assigned to a given taxonomy for each yearly segment and different corpus types are shown. From both perspectives the statistical indicators delivered significant characteristic results for the optimal pre-processed corpus type “n”. With these results a strict distinction between corpora with different pre-processing intensities is only possible on the basis of the statistical indicators introduced. The application of the found indicators can start either from the TRQ correlations or the taxonomy match perspective (see Fig. 113):

- The TRQ correlations always indicate the corpus type.
- If the taxonomy match indicators point out a corpus type “bn” or “b” quality the reason for this behaviour is unclear, until the TRQ correlations are derived. A well matching taxonomy applied to a type “n” corpus must also confirm this corpus type.

A more detailed view of aggregated results with derivatives of action recommendations is shown in Table 43. This may guide knowledge workers to answer the question regarding the “optimal pre-processing level” by statistical

indicators based on TRQ correlations between different vertical corpus segments.

Table 43: TRQ based statistical indicators for corpus quality and action recommendations

Correlation of corpus segment pair C-C_C	Correlation of corpus segment pair C-C_V	Correlation of corpus segment pair C_C-C_V	Conclusion about corpus qualities	Action recommendation
Negative, middle, significant	Positive, strong, significant	Negative, strong, significant	corpus type “n”, optimal corpus size and pre-processing intensity	no actions necessary
~ 0	Positive, middle, significant	Negative, middle, significant	corpus type “n”, sufficient corpus size and pre-processing intensity	increase corpus size (not mandatory)
Positive, middle, significant	Positive, strong, significant	Negative, weak	corpus type “n”, small corpus size, sufficient pre-processing intensity	increase corpus size
Positive, weak	Positive, strong, significant	Negative, weak	corpus type “bn”, even corpus segment size	increase intensity of pre-processing (filter non-target data)
Positive, strong, significant	Positive, strong, significant	Positive, strong, significant	corpus type “b”, volatile corpus segment size	-increase intensity of pre-processing (filter non-target data) -smoothen corpus segment sizes to even yearly length

Based on the given recommendations in Table 43 it is possible to decide whether a given corpus is of sufficient pre-processing intensity or not. The conclusions were drawn from results described in Chapter 4.3.1.1. Another perspective is the distribution analysis of applied taxonomies, which indicates whether the taxonomy applied to the corpus matches the concepts represented in the corpus or not (see Table 44).

Table 44: Indicators for taxonomy match on corpora of different types based on correlations between types and assigned terms per yearly corpus segment

Corpus segment C	Corpus segment C_V	Conclusion about corpus type and the matching quality of taxonomy	Qualitative statement/ Action recommendation
Negative, middle, significant	Negative, strong, significant	corpus type “n”, optimal corpus size, optimal matching taxonomy	no actions necessary
Negative, middle, significant	Negative, middle, significant	corpus type “n”, sufficient corpus size, optimal matching taxonomy	increase corpus size (not mandatory)
Negative, middle, significant	Negative, weak	corpus type “n”, small corpus size, sufficient matching taxonomy	increase corpus size
Positive, middle, significant	Positive, middle significant	corpus type “bn”, not sufficient matching taxonomy	-increase intensity of pre-processing (filter non-target data) -increase quality of taxonomy
Positive, middle, significant	Positive, middle (possibly significant)	corpus type “b”, not sufficient matching taxonomy	-increase intensity of pre-processing (filter non-target data) -smoothen corpus segment sizes -increase quality of taxonomy

A “good” taxonomy cannot by definition perfectly match a corpus with $C_G \neq 0$.

A bad matching rate does not necessarily indicate a bad modelled taxonomy, but can also indicate a non-optimal pre-processing intensity of the used corpus.

With the application of the TMF process and the introduced set of indicators for the quality of pre-processing of corpora a basis for an improvement of TDM results is to be expected. Future research may focus on the following tasks within the data-mining field:

- Defining optimization functions and calculating the optimal level of pre-processing for certain application scenarios

- Analysis of recall and support with corpora of different intensities of pre-processing in concrete TDM scenarios with more sophisticated TDM methods, e.g., SOM

In the computational linguistics field the methods should be evaluated in research on corpus quality, e.g.:

- Applying segmentation on texts without any time information, e.g., to analyse the structure or semantic change within longer texts from a linguistic research perspective
- Measuring the evolution of learner corpora during the learning process of a foreign language and other stylometric tasks

Sources

[Abec04] Abecker, A.; van Elst, L.: Ontologies for Knowledge Management. Hrsg.: Staab,, S.; Studer, R.: Handbook on Ontologies. Berlin, Heidelberg, New York, Springer, 2004, S. 435-454, ISBN 3-540-40834-7

[Alla02a] Allen, J.: Introduction to Topic Detection and Tracking. Hrsg.: Allan, J.: Topic Detection and Tracking: Event-based Information Organization. Massachusetts, Kluwer Academic Publishers, 2002

[Alla02b] Allen, J.; Lavrenko, V.; Swan, R.: Explorations within Topic Tracking and Detection. Hrsg.: Allan, J.: Topic Detection and Tracking: Event-based Information Organization. Massachusetts, Kluwer Academic Publishers, 2002

[Alle05] Allemáo, M. A. F.; Evsukoff, A. G.; Ebecken, N. F. F.: Application of fuzzy models and neural models in financial time series. Hrsg.: Brebbia, C.; Ebecken, N. F. F.; Zanasi, A.: DATA MINING VI - DATA MINING TEXT MINING AND THEIR BUSINESS APPLICATIONS. Southampton, WIT Press, 2005,, S. 475-484, ISBN: 1-84564017-9

[Alli06] Allianz AG: Timeline. Downloaded on 19.04.2006 http://www.allianz.com/azcom/dp/cda/0,,99618-44,00.html

[Anto04a] Antoniou, G.; van Harmelen, F.: A Semantic Web Primer. Massachusetts, MIT Press, 2004, ISBN: 0-262-01210-3

[Anto04b] Antoniou, G.; van Harmelen, F. : Web Ontology Language: OWL. Hrsg.: Staab, S.; Studer, R.: Handbook on Ontologies. Berlin, Heidelberg, New York, Springer, 2004, S.67-92, ISBN 3-540-40834-7

[Atte71] Atteslander, P.: Methoden der empirischen Sozialforschung (Methods of empirical social research). 2. Auflage Berlin, de Gruyter, 1971

[Baad04] Baader, F.; Horrocks, I.; Sattler, U.: Description Logics. Hrsg.: Staab, S.; Studer, R.: Handbook on Ontologies. Berlin, Heidelberg, New York, Springer, 2004, S. 3-28, ISBN 3-540-40834-7

[Baay93] Baayen, H.: Statistical Models for Word Frequency Distributions: A Linguistic Evaluation. Computers and the Humanities. Vol. 26, 1993, pp. 347-363

[Baay98] Baayen, H.; Tweedie, F.: How Variable May a Constant be? Measures of Lexical Richness in Perspective. Computers and the Humanities Vol. 32, 1998, pp. 323-352

[Baay00] Baayen, H.; Tweedie, F. J.: Mixture models for word frequency distributions. Proceedings of JADT 2000: 5es Jourées Internationales d'Analyse Statistique des Données Textuelles. Lausanne, 2000

[Baay02] Baayen, H.; van Halteren, H.; Neijt, A.; Tweedie, F.: An experiment in authorship attribution. Proceedings of JADT 2002: 6es Journées internationales d'Analyse statistique des Données Textuelles. 2002

[Bail78] Bailey, K. D.: Methods of social research. 5. ed. New York, The Free Press, 1978

[Baro03] Baron, S.; Günther, O.; Spiliopoulou, M.: Temporale Aspekte entdeckten Wissens: Ein Überblick über Musterevolution. Proceedings of 2. Workshop GI-AK Knowledge Discovery, 25.2.2003, Universität Leipzig. 2003 , S.55-66

[Bere03] Berendt, B. : Semantische Visualisierung von Webnutzung. Proceedings of 2. Workshop GI-AK Knowledge Discovery, 25.2.2003, Universität Leipzig. 2003 , S. 67-78

[Biss99] Bissantz, N.: Aktive Managementinformation und Data Mining: Neuere Methoden und Ansätze. Hrsg.: Chamoni, P.; Gluchowski, P. : Analytische Informationssysteme: Data Warehouse, On-Line Analytical Processing, Data Mining. 2. Auflage Berlin, Heidelberg, New York, Barcelona, Hongkong, London, Mailand, Paris, Singapur, Tokio, Springer, 1999 ISBN 3-540-65843-2

[Börn03] Börner, K.; Chen, C.; Boyack, K.: Visualizing Knowledge Domains. Hrsg.: Cronin, B.: Annual Review of Information Science & Technology. Band Volume 37. Medford, NJ, Information Today, Inc./American Society for Information Science and Technology, 2003, S.179-255

[Boll05] Trend Analysis of the Digital Library Community. Hrsg.: Bollen, J.; Nelson, M. L.; Manepalli, G.; Nandigam, G.; Manepalli, S.: D-Lib Magazine. Band Volume 11 Number 1. Reston, Corporation for National Research Initiatives, 2005 ISSN 1082-9873

[Borg04] Borgelt, C.; Nürnberger, A.: Visualizing Knowledge Domains. Hrsg.: Abecker, A.; Bickel,, S.; Brefeld, U.; Drost, I.; Henze, N.; Herden, O.; Minor, M.; Scheffer, T.; Stojanovic, L.; Weibelzahl, S.: Proceedings of LWA 2004 Lernen – Wissensentdeckung – Adaptivität Workshopwoche der GI-Fachgruppen/Arbeitskreise. Berlin, Humboldt Universität Berlin, 2004, S.123-130

[Brach04] Brachmann, R. J.; Levesque, H. J.: Knowledge Representation and Reasoning. San Francisco, Elsevier, 2004, ISBN: 1-55860-932-6

[Brad05] Bradley, J.: Documents and Data: Modelling Materials for Humanities Research In XML and Relational Databases, Literary & Linguistic Computing – Journal of the Association for Literary and Linguistic Computing and the Association for Computers and the Humanities, 2005, *Vol. 20 Number 1,* London, Oxford, p. 133-151, ISSN: 0268-1145

[Bras04] Brase, J.; Nejdl, W.: Ontologies and Metadata for eLearning, Hrsg: Staab, S.; Studer, R.: Handbook on Ontologies, Berlin, Heidelberg, New York, Springer, 2004, S. 555-573, ISBN 3-540-40834-7

[Brui02] Bruijn, de B.; Martin, J.: Literature mining in molecular biology, Institute for Information Technology, national Research Council, Canada, 2002

[Cabe98] Cabena, P.; Hadjinian, P.; Stadler, R.; Verhees, J.; Zanasi, A.: Discovering data mining - from concept to implementation, Upper Saddle River, 1998

[Capt05] Captain's Universe, downloaded on 10/21/2005, URL: http://www.captain.at/review-internet.php

[Cast04] Castellanos, M.: HotMiner: Discovering Hot Topics from Dirty Text, Hrsg.: Berry, M.: Survey of Text Mining: Clustering, Classification and Retrieval, New York, Springer, 2004

[Cham99a] Chamoni, P.; Gluchowski, P.: Analytische Informationssysteme - Einordnung und Überblick. Hrsg.: Chamoni, P.; Gluchowski, P.: Analytische Informationssysteme: Data Warehouse, On-Line Analytical Processing, Data Mining, 2. Auflage Berlin, Heidelberg, New York, Barcelona, Hongkong, London, Mailand, Paris, Singapur, Tokio, Springer, 1999, ISBN 3-540-65843-2

[Cham99b] Chamoni, P.; Gluchowski, P.: Entwicklungslinien und Architekturkonzepte des On-Line Analytical Processing. Hrsg.: Chamoni, P.; Gluchowski, P.: Analytische Informationssysteme: Data Warehouse, On-Line Analytical Processing, Data Mining, 2. Auflage Berlin, Heidelberg, New York, Barcelona, Hongkong, London, Mailand, Paris, Singapur, Tokio, Springer, 1999, ISBN 3-540-65843-2

[Cham99c] Chamoni, P.: Ausgewählte Verfahren des Data Mining Hrsg.: Chamoni, P.; Gluchowski, P.: Analytische Informationssysteme: Data Warehouse, On-Line Analytical Processing, Data Mining, 2. Auflage Berlin, Heidelberg, New York, Barcelona, Hongkong, London, Mailand, Paris, Singapur, Tokio, Springer, 1999, ISBN 3-540-65843-2

[Cham00] Chamoni, P.: On-Line Analytical Processing (OLAP), Hrsg.: Hippner, H.; Küsters, U.; Meyer, M.; Wilde, K.D.: Handbuch Data Mining im Marketing, Braunschweig,Vieweg, 2000

[Chen04] Chen, C.: Information Visualization – Beyond the Horizon, Second Edition, London, Springer, 2004, ISBN: 1852337893

[Cimi03] Cimiano, P.; Staab, S.; Tane, J.: Automatic acquisition of taxonomies from text: FCA meets NLP. Proceedings of the International Workshop on Adaptive Text Extraction and Mining, downloaded on 9/25/2009, URL: http://www.dcs.shef.ac.uk/~fabio/ATEM03/cimiano-ecml03-atem.pdf

[Codd93] Codd, E.F.; Codd, S.B.; Sally, C.T.: Providing OLAP (on-line analytical processing) to user-analysts – an IT mandat. White Paper, E.F. Codd & Associates, 1993

[Corn04] Cornelson, M.; Greengrass, E.; Grossmann, R. L.; Karidi, R.; Shnidman, D.: Combining Families of Information Retrieval Algorithms Using Metalearning. Hrsg.: Berry, M.: Survey of Text Mining: Clustering, Classification and Retrieval, New York, Springer, 2004

[Crou90] Crouch, C. J.: An approach to the automatic construction of global thesauri. Information Processing and Management, 1990, *26,* p. 629-640

[Curi05] Curia, R.; Ettorre, M.; Gallucci, L.; Tiritan,, S.; Rullo, P.: Textual document pre-processing and feature extraction in OLEX. Hrsg.: Brebbia, C.; Ebecken, N. F. F.; Zanasi, A.: DATA MINING VI - DATA MINING TEXT MINING AND THEIR BUSINESS APPLICATIONS, Southampton, WIT Press, 2005, 163-173, ISBN: 1-84564017-9

[Dame05] Damerau, J; Indurkhya, N.; Weiss, S. M.; Zhang, T.: Text Mining – Predictive Methods Analyzing Unstructured information. New York, Springer, Science+Business Media Inc., 2005

[Dege99c] Degen, H.: Statistische Methoden zur visuellen Exploration mehrdimensionaler Daten. Hrsg.: Chamoni, P.; Gluchowski, P.: Analytische Informationssysteme: Data Warehouse, On-Line Analytical Processing, Data Mining. 2. Auflage, Berlin, Heidelberg, New York, Barcelona, Hongkong, London, Mailand, Paris, Singapur, Tokio, Springer, 1999, ISBN 3-540-65843-2

[Desj05] Desjardins;G.; Godin, R.; Proul, R.: A genetic algorithm for text mining. Hrsg.: Brebbia, C.; Ebecken, N. F. F.; Zanasi, A.: DATA MINING VI - DATA MINING TEXT MINING AND THEIR BUSINESS APPLICATIONS, Southampton, WIT Press, 2005, 133-142, ISBN: 1-84564017-9

[Doan04] Doan, A.; Madhavan, J.; Dorningos, P.; Halevy, A.: Ontology Matching: A Machine Learning Approach. Hrsg.: Staab, S.; Studer, R.:

Handbook on Ontologies, Berlin, Heidelberg, New York, Springer, 2004, 385-403, ISBN 3-540-40834-7

[Doan05] Doan, S.; Horiguchi, S.: A multi-criteria decision making approach in feature selection for enhancing text categorization. Hrsg.: Brebbia, C.; Ebecken, N. F. F.; Zanasi, A.: DATA MINING VI - DATA MINING TEXT MINING AND THEIR BUSINESS APPLICATIONS, Southampton, WIT Press, 2005, 77-87, ISBN: 1-84564017-9

[Dörr99] Dörre, J.; Gerstl, P.; Seiffert, R.: Text Mining: Finding Nuggets in Mountains of Textual Data. Proceedings of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Diego, 1999

[Dörr00] Dörre, J.; Gerstl, P.; Seiffert, R.: Text Mining. Hrsg.: Hippner, H.; Küsters, U.; Meyer, M.; Wilde, K.D.: Handbuch Data Mining im Marketing, Braunschweig, Vieweg, 2000

[Druc70] Drucker, P. F.: The Age of Discontinuity - Guidelines to our Changing Society, Orbit, 1970

[Düsi99] Düsing, J.: Knowledge Discovery in Databases und Data Mining. Hrsg.: Chamoni, P.; Gluchowski, P. : Analytische Informationssysteme: Data Warehouse, On-Line Analytical Processing, Data Mining. 2. Auflage, Berlin, Heidelberg, New York, Barcelona, Hongkong, London, Mailand, Paris, Singapur, Tokio, Springer, 1999, ISBN 3-540-65843-2

[Eklu04] Eklund, P.; Cole, R.; Roberts, N.: Retrieving and Exploring Ontology-based Information. Hrsg.: Staab, S.; Studer, R.: Handbook on Ontologies, Berlin, Heidelberg, New York, Springer, 2004, 405-414, ISBN 3-540-40834-7

[Eppl04] Eppler, M. J. ; Burkhard, R. A.: Knowledge Visualization Towards a New Discipline and its Fields of Application, Schweiz, 2004

[Espí05] Espindola, R. P.; Ebecken, N. F. F.On extending F-measure and G-mean metrics to multi-class problems. Hrsg.: Brebbia, C.; Ebecken, N. F. F.; Zanasi, A.: DATA MINING VI - DATA MINING TEXT MINING

AND THEIR BUSINESS APPLICATIONS, Southampton, WIT Press, 2005, ISBN: 1-84564017-9

[Faul05] Faulstich, L.; Leser, U.; Lüdeling, A.: Storing und Querying Historical Texts in a Relational Database, Berlin, 2005

[Fayy96] Fayyad, U.M.; Piatetsky-Shapiro, G.; Smyth, P.: From data mining to knowledge discovery: an overview. Hrsg.: Fayyad, U.M.; Piatetsky-Shapiro, G.; Smyth, P.; Uthurusamy, R.: Advances in knowledge discovery and data mining, Menlo Park (California), 1996, p. 1-34

[Ferr05] Ferrari, M.: ROI in text mining projects. Hrsg.: Zanasi, A.: Text Mining und its Applications to Intelligence, CRM und Knowledge Management, Series: Advances in Management Information, Vol. 2, WIT Press, 2005, 155-184, ISBN: 1-85312-995-X

[Fisc02] Fiscus, J., G; Doddington, G., R.: Topic Detection and Tracking Evaluation Overview. Hrsg.: Allen, J., Topic Detection and Tracking: Event-based Information Organization, Massachusetts, Kluwer Academic Publishers, 2002

[Flui04] Fluit, C.; Sabou, M.; van Harmelen, F.: Supporting User Tasks through Visualisation of Light-weight Ontologies. Hrsg.: Staab, S.; Studer, R.: Handbook on Ontologies, Berlin, Heidelberg, New York, Springer, 2004, 415-432, ISBN 3-540-40834-7

[Frig04] Frigui, H.; Nasraoui, O.: Simultaneous Clustering and Dynamic Keyword Weighting for Text Documents. Hrsg.: Berry, M.: Survey of Text Mining: Clustering, Classification and Retrieval, New York, Springer, 2004

[Gärd04] Gärdenfors, P.: Conceptual Spaces – The Geometry of Thought. London, MIT Press, 2004, ISBN: 0-262-07199-1

[Gart03] Gartner: Understanding Gartner's Hype Cycles, Stamford, Gartner, 2003

[Gome04a] Gómez-Pérez, A.: Ontology Evaluation. Hrsg: Staab, S.; Studer, R.: Handbook on Ontologies, Berlin, Heidelberg, New York, Springer, 2004 251-273, ISBN: 3-540-40834-7

[Gome04b] Gómez-Pérez, A.; Fernández-López, M.; Corcho, O.: Ontological Engineering, London, Springer, 2004, ISBN: 1-85233-551-3

[Grah00] Graham, N.: AUTOMATIC DETECTION OF AUTHORSHIP CHANGES WITHIN SINGLE DOCUMENTS. Toronto Graduate Department of Computer Science, 2000

[Grob03] Grobelnik, M.: Ljubljana Institut Jozef Stefan, downloaded on 01/05/2007, URL: http://ai.ijs.si/Mezi/pedagosko/AU4DataMining.ppt

[Grub93] Gruber, T. R.: Toward Principles for the Design of Ontologies Used for Knowledge Sharing. Formal Ontology in Conceptual Analysis and Knowledge Representation, Kluwer Academic Publishers, 1993

[Hahn04] Hahn, U. ; Schulz, S.: Building a Very Large Ontology from Medical Thesauri. Hrsg.: Staab, S.; Studer, R.: Handbook on Ontologies,Berlin, Heidelberg, New York,Springer,2004, 133-150, ISBN 3-540-40834-7

[Havr02] Havre, S.; Hetzler, E.; Whitney, P.; Nowell, L.: ThemeRiver: Visualizing thematic changes in large document collections. IEEE Transactions on Visualization and Computer Graphics, 2002, *8(1)*, Jan – Mar 2002

[Hear99] Hearst, M. A.: Untangling Text Mining. Proceedings of ACL'99: the 37th Annual Meeting of the Association for Computational Linguistics, Maryland, 1999, downloaded on 06/05/2004, URL: http://people.ischool.berkeley.edu/~hearst/papers/acl99/acl99-tdm.html

[Herd60] Herdan, G.: Type-Token Mathematics. The Hague, Mourton & Co., 1960

[Hida02] Hidalgo, J. M. G.: Text Mining and Internet Content Filtering. Departamento de Inteligencia Artificial Universidad Europea, CEES, 2002

[Hinr02] Hinrichs, H.: Datenqualitätsmanagement in Data Warehouse-Systemen. Universität Oldenburg, Dissertation, 2002

[Hirs04] Hirst, G.: Ontology and the Lexicon. Hrsg.: Staab, S.; Studer, R.: Handbook on Ontologies, Berlin, Heidelberg, New York, Springer, 2004, 209-229, ISBN 3-540-40834-7

[Hoov03] Hoover, L. D.: Another Perspective on Vocabulary Richness. Computers and the Humanities. Vol. 37, Massachusetts, Kluwer Academic Publishers, 2003, p. 151-178

[Howl04] Howland, P.; Park, H.: Cluster-Preserving Dimension Reduction Methods for Efficient Classification of Text Data. Hrsg.: Berry, M.: Survey of Text Mining: Clustering, Classification and Retrieval, New York, Springer, 2004

[Kahl93] Kahlert, J.: Fuzzy-Logik und Fuzzy-Control eine anwendungsorientierte Einführung mit Begleitsoftware. Braunschweig, Wiesbaden, Friede Vieweg & Sohn Verlagsgesellschaft mbH, 1993, ISBN 3-528-05304-6

[Kall03] Kalledat, T.: Separation of long-term constant elements in the field of information technology from short existing trends based on unstructured data. Hrsg.: Viehweger, B.: Perspectives in Business Informatics Research, Proceedings of the BIR-2003-Conference, Aachen, Shaker Verlag, 2003, S. 167-183

[Kall04] Kalledat, T.: Perspektiven der Nutzung von Data-Mining-Technologien in der Energieversorgungswirtschaft. ew-Elektrizitätswirtschaft – Das Magazin für die Energiewirtschaft, 2004, *7*, S. 48-53, ISSN: 1619-5795-D9785D

[Kall05] Kalledat, T.: Meta Data based pre-filtering for large text collections in KD Processes. Proceedings of the BIR-2005-Conference, Skövde, 2005

[Kall06] Kalledat, T.: Evaluation of Corpus Measures for the use in Data Mining scenarios. Proceedings of CaSTA 2006: Breadth of Text - A Joint Computer Science and Humanities Computing Conference. New

Brunswick, University of New Brunswick, 2006, URL: http://www.lib.unb.ca/casta2006/viewabstract.php?id=412006

[Kara05] Karanikolas, N. N.; Skourlas, C.: Valve rule induction for text classification based on key-terms. Hrsg.: Brebbia, C.; Ebecken, N. F. F.; Zanasi, A.: DATA MINING VI - DATA MINING TEXT MINING AND THEIR BUSINESS APPLICATIONS, Southampton, WIT Press, 2005, 175-181, ISBN: 1-84564017-9

[Khor05] Khordad, M.; Shamsfard, M.; Kazemeyni, F.: A hybrid method to categorize HTML documents. Hrsg.: Brebbia, C.; Ebecken, N. F. F.; Zanasi, A.: DATA MINING VI - DATA MINING TEXT MINING AND THEIR BUSINESS APPLICATIONS, Southampton, WIT Press, 2005, 331-340, ISBN: 1-84564017-9

[Klir95] Klir, G. J.; Yuan, B.: FUZZY SETS AND FUZZY LOGIC – Theory and Appilcations, New Jersey, Prentice Hall, 1995, ISBN: 0-13-101171-5

[Koba04] Kobayashi, M.; Aono, M.: Vector Space Models for Search and Cluster Mining. Hrsg.: Berry, M.: Survey of Text Mining: Clustering, Classification and Retrieval. New York, Springer, 2004

[Kont04] Kontostathis, A.; Galitsky, L.; Pottenger, W.; Roy, S.; Phelps, D.: A Survey of Emerging Trend Detection in Textual Data Mining. Hrsg.: Berry, M.: Survey of Text Mining: Clustering, Classification and Retrieval, New Yor, Springer, 2004

[Kühn02] Kühnlein, C.; Karlsson, M.; Klenner, M.: Bewertung ausgewählter Systeme zum Text Mining in Fortbildungsseminar Text Mining, Institut für Computerlinguistik, Universität Zürich, 16. Oktober 2002, downloaded on 07/05/2003, URL: http://www.ifi.unizh.ch/cl/FoSI02/tm-3.mk.eval.pdf

[Küpp99] Küppers, B.: Data Mining in der Praxis - ein Ansatz zur Nutzung der Potentiale von Data Mining im betrieblichen Umfeld. Frankfurt/Main, 1998

[Küst00] Küsters, U.: Data Mining Methoden: Einordnung und Überblick. Hrsg.: Hippner, H.; Küsters, U.; Meyer, M.; Wilde, K.D.: Handbuch Data Mining im Marketing, Braunschweig, Vieweg, 2000

[Kwia05] Kwiatkowskal, M.; Ayas, N. T.; Ryan, F.: Evaluation of clinical prediction rules using a convergence of knowledge-driven and data-driven methods: a semio-fuzzy approach. Hrsg.: Brebbia, C.; Ebecken, N. F. F.; Zanasi, A.: DATA MINING VI - DATA MINING TEXT MINING AND THEIR BUSINESS APPLICATIONS, Southampton, WIT Press, 2005, 411-420, ISBN: 1-84564017-9

[Larr05] Larreina, S.; Hernando, S.: Application of technology prospective to business sectorial studies. Hrsg.: Brebbia, C.; Ebecken, N. F. F.; Zanasi, A.: DATA MINING VI - DATA MINING TEXT MINING AND THEIR BUSINESS APPLICATIONS, Southampton, WIT Press, 2005, 345-352, ISBN: 1-84564017-9

[Lavr00] Lavrenko, V.; Schmill, M.; Lawrie, D.; Ogilvie, P.; Jensen, D.; Allan, J.: Mining of Concurrent Text and Time Series. Proceedings of KDD 2000 Conference, 2000, 37-44

[Lebe05b] Lebeth, K.; Lorenz, M.; Störl, U.: Text mining based knowledge management in banking. Hrsg.: Zanasi, A.: Text Mining und its Applications to Intelligence, CRM und Knowledge Management. Series: Advances in Management Information, Vol. 2, WIT Press, 2005, 271-278, ISBN: 1-85312-995-X

[Leek02] Leek, T.; Schwartz, R.; Srinivasa, S.: Probabilistic Approaches To Topic Detection and Tracking. Hrsg.: Allen, J.: Topic Detection and Tracking: Event-based Information Organization, Massachusetts, Kluwer Academic Publishers, 2002

[Lend98] Lenders, W.; Willée, G.: Linguistische Datenverarbeitung. 2. Auflage, Opladen/ Wiesbaden, Westdeutscher Verlag, 1998

[Maed04] Maedche, A.; Staab, S.: Ontology Learning. Hrsg.: Staab, S.; Studer, R.: Handbook on Ontologies, Berlin, Heidelberg, New York, Springer, 2004, 173-190, ISBN 3-540-40834-7

[Mand05] Mandreoli, F.; Martoglia, R.; Tiberio, P.: Text clustering as a mining task. Hrsg.: Zanasi, A.: Text Mining und its Applications to Intelligence, CRM und Knowledge Management. Series: Advances in Management Information, Vol. 2, WIT Press, 2005, 76-108, ISBN: 1-85312-995-X

[McBr04] McBride, B. The Resource Description Framework (RDF) and its Vocabulary Description Language RDF. Hrsg.: Staab, S.; Studer, R.: Handbook on Ontologies, Berlin, Heidelberg, New York, Springer, 2004, 51-66, ISBN, 3-540-40834-7

[Mend99] Mendonca, M.; Sunderhaft, N. L.: Mining Software Engineering Data: A Survey - A DACS State-of-the-Art Report, Rome DoD Data & Analysis Center for Software (DACS), 1999

[Mert95] Mertens, P.: Von den Moden zum Trend. Hrsg.: König, W.: Wirtschaftsinformatik '95, Wettbewerbsfähigkeit - Innovation – Wirtschaftlichkeit, Heidelberg, 1995, 25-65

[Meye02] Meyer, Dr. M.: Einsatz von Klassifikation und Prognose im Web Mining. Hrsg.: Hippner, H.; Merzenich, M.; Wilde, K., D.: Handbuch Web Mining im Marketing - Konzepte, Systeme, Fallstudien, 1. Auflage Braunschweig/ Wiesbaden, Vieweg, 2002, ISBN 3-528-05794-7

[Mian05] Miangah, T. M.; Khalafi, A. D.: Word Sense Disambiguation Using Target Language Corpus in a Machine Translation System. Literary & Linguistic Computing – journal of the Association for Literary and Linguistic Computing and the Association for Computers and the Humanities, 2005, *Vol. 20 Number 2*, London, Oxford, 237-249, ISSN: 0268-1145

[Mili05] Milic-Frayling, N.: Text processing und information retrieval. Hrsg.: Zanasi, A.: Text Mining und its Applications to Intelligence, CRM und Knowledge Management. Series: Advances in Management Information, Vol. 2, WIT Press, 2005, 1-45, ISBN: 1-85312-995-X

[Missi04] Missikoff, M.; Taglino, F.: An Ontology-based Platform for Semantic Interoperability. Hrsg.: Staab, S.; Studer, R.: Handbook on Ontologies,

Berlin, Heidelberg, New York, Springer, 2004, 617-633, ISBN 3-540-40834-7

[Mlad 05] Mladenic, D.; Grobelnik, M.: Text categorization. Hrsg.: Zanasi, A.: Text Mining und its Applications to Intelligence, CRM und Knowledge Management. Series: Advances in Management Information, Vol. 2, WIT Press, 2005, 132-143, ISBN: 1-85312-995-X

[Moen00] Moens, M.-F.: Automatic Indexing and Abstracting of Document Texts. Massachusetts, Kluwer, Academic, Publishers, 2000

[Mont99] Montes-y-Gómez, M. ; Gelbukh, A. F. ; López-López, A.: Text Mining as a Social Thermometer. Proc. Text Mining workshop at 16th International Joint Conference on Artificial Intelligence (IJCAI'99), Stockholm, 1999, 103-107

[Nahm01] Nahm, U., Y.: A Roadmap to Text Mining and Web Mining. University of Austin, downloaded on 02/19/2005 URL: http://www.cs.utexas.edu/users/pebronia/text-mining/

[Nefi96] Nefiodow, L. A.: Der sechste Kondratieff: Wege zur Produktivität und Vollbeschäftigung im Zeitalter der Information, Sankt Augustin, Rhein-Sieg Verlag, 1996

[Noy04] Noy, N. F.: Tools for Mapping and Merging Ontologies. Hrsg.: Staab, S.; Studer, R.: Handbook on Ontologies, Berlin, Heidelberg, New York, Springer, 2004, 365-384, ISBN 3-540-40834-7

[Obe04] Oberle, D.; Spyns, P.: The Knowledge Portal “OntoWeb”. Hrsg.: Staab, S.; Studer, R.: Handbook on Ontologies, Berlin, Heidelberg, New York, Springer, 2004, 499-516, ISBN 3-540-40834-7

[Orlo83] Orlov, J. K.: Ein Modell der Häufigkeitsstruktur des Vokabulars. Hrsg.: Guiter, H.; Arapov, M.: Studies on Zipfs Law, Bochum, Brockmeyer, 1983

[Otte04] Otte, R.; Otte, V.; Kaiser, V.: Data Mining für die industrielle Praxis, München, Wien, Carl Hanser, 2004

[Paaß04] Paaß, G.; Kindermann, J.; Leopold, E.: Learning Prototype Ontologies by Hierachical Latent Semantic Analysis. Hrsg.: Abecker, A.; Bickel, S.; Brefeld, U.; Drost, I.; Henze, N.; Herden, O.; Minor, M.; Scheffer, T.; Stojanovic, L.; Weibelzahl, S.: Proceedings of LWA 2004 Lernen – Wissensentdeckung – Adaptivität Workshopwoche der GI-Fachgruppen/Arbeitskreise, Humboldt Universität, Berlin, 2004, 193-205

[Pan05] Pan, J., Z.; Stamou, G.; Tzouvaras, V.; Horrocks, I.: f-SWRL: A Fuzzy Extension of SWRL, 2005

[Pazi05] Pazienza, M. T.: Information extraction and …. surroundings. Hrsg.: Zanasi, A.: Text Mining und its Applications to Intelligence, CRM und Knowledge Management, Series: Advances in Management Information, Vol. 2, WIT Press, 2005, 47-74, ISBN: 1-85312-995-X

[Pete05b] Peters, G.: Media industry: how to improve documentalists efficiency. Hrsg.: Zanasi, A.: Text Mining und its Applications to Intelligence, CRM und Knowledge Management, Series: Advances in Management Information, Vol. 2, WIT Press, 2005, 293-298, ISBN: 1-85312-995-X

[Pohl04] Pohle, C.: Integrating Domain Knowledge for Data Mining Post-Processing. Hrsg.: Abecker, A.; Bickel,, S.; Brefeld, U.; Drost, I.; Henze, N.; Herden, O.; Minor, M.; Scheffer, T.; Stojanovic, L.; Weibelzahl, S.: Proceedings of LWA 2004 Lernen – Wissensentdeckung – Adaptivität Workshopwoche der GI-Fachgruppen/Arbeitskreise, Humboldt Universität, Berlin, 2004, 76-83

[Poli05] Politi, A.: A critical appraisal of text mining in an intelligence environment. Hrsg.: Zanasi, A.: Text Mining und its Applications to Intelligence, CRM und Knowledge Management. Series: Advances in Management Information, Vol. 2, WIT Press, 2005, 209-217, ISBN: 1-85312-995-X

[Raja01] Rajaraman, K.; Tan, A.: Topic Detection, Tracking and Trend Analysis Using Self-organizing Neural Networks, Singapore, Kent Ridge Digital Labs, 2001, 102-107

[Raub05] Rauber, A.; Merkl, D.: Mining Text Archives: Creating Readable Maps to Structure and Describe Document Collections, Institute of Software Technology, University of Technology, Vienna, 2005

[Rehb04] Rehbein, I.: Eine quantitative Untersuchung der Lexik von Zeitungstexten vor und nach dem 11. September. Berlin, Magisterarbeit am Institut für Korpuslinguistik der Humboldt Universität zu Berlin, 2004

[Roma05] Romanov, V.; Pantileeva, E.: Knowledge discovery in large text databases using the MST algorithm. Hrsg.: Brebbia, C.; Ebecken, N. F. F.; Zanasi, A.: DATA MINING VI - DATA MINING TEXT MINING AND THEIR BUSINESS APPLICATIONS, Southampton, WIT Press, 2005, 153-162, ISBN: 1-84564017-9

[Sala05] Salazar, A.; Gosalbez, J.; Bosch, I.: Mining association rules from qualitative and quantitative clustering. Hrsg.: Brebbia, C.; Ebecken, N. F. F.; Zanasi, A.: DATA MINING VI - DATA MINING TEXT MINING AND THEIR BUSINESS APPLICATIONS, Southampton, WIT Press, 2005, 299-310, ISBN: 1-84564017-9

[Sant05] Santos, M. F.; Cortez, P.; Quintelal, H.; Pinto, F.: A clustering approach for knowledge discovery in database marketing. Hrsg.: Brebbia, C.; Ebecken, N. F. F.; Zanasi, A.: DATA MINING VI - DATA MINING TEXT MINING AND THEIR BUSINESS APPLICATIONS, Southampton, WIT Press, 2005, 399-407, ISBN: 1-84564017-9

[Sche99] Schelp, J.: Konzeptionelle Modellierung mehrdimensionaler Datenstrukturen. Hrsg.: Chamoni, P.; Gluchowski, P. : Analytische Informationssysteme: Data Warehouse, On-Line Analytical Processing, Data Mining. 2. Auflage, Berlin, Heidelberg, New York, Barcelona, Hongkong, London, Mailand, Paris, Singapur, Tokio, Springer, 1999, ISBN 3-540-65843-2

[Seba05] Sebastiani, F.: Text categorization. Hrsg.: Zanasi, A.: Text Mining und its Applications to Intelligence, CRM und Knowledge Management. Series: Advances in Management Information, Vol. 2, WIT Press, 2005, 109-129, ISBN: 1-85312-995-X

[Sene04] Senellart, P.; Blondel, V. D.: Automatic Discovery of Similar Words. Berry, M.: Survey of Text Mining: Clustering, Classification and Retrieval. New York, Springer, 2004

[Smit91] Smith, G. W.: Computers and Human Language, New York, Oxford, Oxford University Press,1991

[Smit02] Smith, D. A.: Detecting and Browsing Events in Unstructured Text, Perseus Project, Tufts University, Medford, 2002

[Spil02] Spiliopoulou, M.; Winkler, K.: Text Mining auf Handelsregistereinträgen: Der SAS Enterprise Miner im Einsatz. Hrsg.: Wilde, K. D.; Hippner, H.; Merzenich, M.: Data Mining: Mehr Gewinn aus Ihren Kundendaten. Düsseldorf, Verlagsgruppe Handelsblatt, S. 117-124

[Stam01] Stamatatos, E.; Fakotakis, N.; Kokkinakis, G.: Computer-Based Authorship Attribution Without Lexical Measures. Computers and the Humanities, 35, Netherlands, Kluwer Academic Publishers, 2001, 193-214

[Stei00] Steiger, C.: WISSENSMANAGEMENT IN BERATUNGSPROJEKTEN AUF BASIS INNOVATIVER INFORMATIONS- UND KOMMUNIKATIONSTECHNOLOGIEN: DAS SYSTEM K3. Dissertation, Paderborn, Universität-Gesamthochschule-Paderborn, 2000

[Stra05] Straccia, U.: Fuzzy Description Logics with Concrete Domains. Technical Report, 2005, TR-03PisaISTI-CNR2005

[Stud98] Studer, R.; Benjamins, V.R.; Fensel, D.: Knowledge Engineering: Principles and Methods. Data & Knowledge Engineering, 25(1-2), 1998, 161-197

[Sull05] Sullivan, D.: Application integration in applied text mining. Hrsg.: Zanasi, A.: Text Mining und its Applications to Intelligence, CRM und Knowledge Management. Series: Advances in Management Information, Vol. 2, WIT Press, 2005, 145-154, ISBN: 1-85312-995-X

[Sure04] Sure, Y. ; Staab, S. ; Studer, R.: On-To-Knowledge Methodology (OTKM). Hrsg.: Staab, S.; Studer, R.: Handbook on Ontologies, Berlin, Heidelberg, New York, Springer, 2004, 117-132, ISBN 3-540-40834-7

[Tagu04] Tague, N., R.: The Quality Toolbox, Second Edition, ASQ Quality Press, 2004, 390-392

[Tan99] Tan, A.-H.: Text Mining: The State of the Art and the Challenges. Proceedings of the PAKDD 1999 Workshop on Knowledge Discovery from Advanced Databases. Peking, 1999, S. 65-70

[Vafo05] Vafopoulos, M.; Aggelis, V.; Platis, A.: HyperClustering: from the digital divide to a GRID e-workspace. Hrsg.: Brebbia, C.; Ebecken, N. F. F.; Zanasi, A.: DATA MINING VI - DATA MINING TEXT MINING AND THEIR BUSINESS APPLICATIONS, Southampton, WIT Press, 2005, 311-320, ISBN: 1-84564017-9

[Vall05] Vallet, D.; Fernández, M.; Castells, P.: An Ontology-Based Information Retrieval Model. Book Series Lecture Notes in Computer Science. Berlin/ Heidelberg, Springer, 2005, 455-470, ISBN 978-3-540-26124-7

[Wang04] Wang, P.; Bownas, J., Berry, M.: Trend and Behavior Detection from Web Queries. Hrsg.: Berry, M.: Survey of Text Mining: Clustering, Classification and Retrieval, New York, Springer, 2004

[Wiki06] Wikipedia the free encyclopedia: Text. Wikipedia.org, downloaded on 03.04.2006, URL: http://en.wikipedia.org/wiki/Text

[Witt05] Witt, A.; Goecke, D.; Sasaki, F.; Lüngen, J.: Unification of XML Documents with Concurrent Markup. Literary & Linguistic Computing – journal of the Association for Literary and Linguistic Computing and the Association for Computers and the Humanities. 2005, *Vol. 20 Number 1*, London, Oxford, 103-116, ISSN: 0268-1145

[Yamr02] Yamron, J. P.; Gillick, L.; van Mulbregt, P.: Statistical Models of Topical Content. Hrsg.: Allen, J.: Topic Detection and Tracking: Event-based Information Organization. Massachusetts, Kluwer Academic Publishers, 2002

[Yang99] Yang, Y.; Carbonell, J.; Brown, R.; Pierce, T.; Archibald, B. T.; Liu, X.: Learning approaches for Detecting and tracking news events. Pittsburgh, Language Technology Institute, Carnegy Mellon University, 1999

[Yang02] Yang, Y.; Carbonell, J.; Brown, R.; Lafferty, J.; Pierce, T.; Ault, T.: Multi-strategy Learning for Topic Detection and Tracking. Hrsg.: Allen, J.: Topic Detection and Tracking: Event-based Information Organization. Massachusetts, Kluwer Academic Publishers, 2002

[Yule44] Yule, G. U.: The statistical study of literary vocabulary. Cambridge University Press, 1944

[Zade65] Zadeh, L.: Fuzzy Sets. In Control. 1965, *Vol. 8*, 338-353

[Zade99] Zadeh, L.: From Computing with Numbers to Computing with Words - From Manipulation of Measurements to Manipulation of Perceptions. IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS - I: FUNDAMENTAL THEORY AND APPLICATIONS. 1999, *Vol. 45*, 105-119, ISSN 1057-7122(99)00546-2

[Zade02] Zadeh, L.: Toward a perception-based theory of probabilistic reasoning with imprecise probabilities. Journal of Statistical Planning and Inference. 2002, *Vol. 105*, Elsevier Science B.V., 233-264, ISSN: 50378-3758(01)00212-9

[Zana05a] Zanasi, A.: Text mining: a new technology paradigm? Hrsg.: Zanasi, A.: Text Mining und its Applications to Intelligence, CRM und Knowledge Management. Series: Advances in Management Information, Vol. 2, WIT Press, 2005, ISBN: 1-85312-995-X

[Zana05b] Zanasi, A.: Open sources automatic analysis for corporate and government intelligence. Hrsg.: Zanasi, A.: Text Mining und its Applications to Intelligence, CRM und Knowledge Management. Series: Ad-

vances in Management Information, Vol. 2, WIT Press, 2005, 185-208, ISBN: 1-85312-995-X

[Zana05c] Zanasi, A.: Text mining tools. Hrsg.: Zanasi, A.: Text Mining und its Applications to Intelligence, CRM und Knowledge Management. Series: Advances in Management Information, Vol. 2, WIT Press, 2005, 315-327, ISBN: 1-85312-995-X

[Zipf49] Zipf, G. K.: Human Behavior and The Principle of Least Effort: An Introduction to Human Ecology (Addison–Wesley, Cambridge, MA); reprinted in Zipf, G. K. (1972) Human Behaviour and the Principle of Least Effort: An Introduction to Human Ecology (Hafner, New York), 19–55

Appendix

Table 45: Corpus data sets of descriptive statistics of CW_{5k}

Descriptive Statistics

	N	Range	Minimum	Maximum	Sum	Mean	Std.	Variance	Skewness		Kurtosis	
	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Std. Error
Cw5kCTer	29	597	499726	500323	14502249	500077,55	148,407	22024,613	-,544	,434	-,393	,845
Cw5kbCTer	29	17078749	10745195	27823944	4,3E+08	1,5E+07	4254985	1,8E+13	1,835	,434	2,909	,845
Cw5kbuCTer	29	17004713	10676175	27680888	4,3E+08	1,5E+07	4237624	1,8E+13	1,825	,434	2,875	,845
Cw5kbunCTer	29	535	499784	500319	14499822	499993,86	98,006	9605,195	1,004	,434	4,716	,845
Cw5kbun2CTer	29	247	499819	500066	14499871	499995,55	43,332	1877,685	-2,391	,434	9,771	,845
Valid N (listwise)	29											

Table 46: Corpus data sets of descriptive statistics off CW_{1k}

Descriptive Statistics

	N	Range	Minimum	Maximum	Sum	Mean		Std.	Variance	Skewness		Kurtosis	
	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Statistic	Statistic	Std. Error	Statistic	Std. Error
Cw1kCTer	29	358	99811	100169	2901069	100036,86	18,542	99,852	9970,409	-,734	,434	-,615	,845
Valid N (listwise)	29												

Table 47: Corpus data sets of descriptive statistics of $Al1k_{S1}$ and $Al1k_{S2}$

Descriptive Statistics

	N	Range	Minimum	Maximum	Sum	Mean		Std.	Variance	Skewness		Kurtosis	
	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Statistic	Statistic	Std. Error	Statistic	Std. Error
Al100S1CTer	39	23	986	1009	38956	998,87	,907	5,662	32,062	-,248	,378	-,560	,741
Al100S2CTer	39	19	989	1008	38921	997,97	,921	5,751	33,078	,478	,378	-1,059	,741
Valid N (listwise)	39												

Table 48: Corpus data sets of descriptive statistics of CW_{5k} corpus segments based on TRQ measure

Descriptive Statistics

	N	Range	Minimum	Maximum	Sum	Mean		Std.	Variance	Skewness		Kurtosis	
	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Statistic	Statistic	Std. Error	Statistic	Std. Error
Cw5kCRepQuo	29	1,42354	7,88483	9,30837	245,76323	8,4745940	,05321350	,28656347	,082	,871	,434	1,634	,845
Cw5kCcRepQuo	29	2,61862	61,13767	63,75629	1815,164	62,59188	,15668346	,84376624	,712	-,266	,434	-1,431	,845
Cw5kCvRepQuo	29	,50952	1,95259	2,46211	61,71642	2,1281526	,03039049	,16365778	,027	,596	,434	-1,214	,845
Valid N (listwise)	29												

Table 49: Corpus data sets descriptive statistics of CW_{1k} corpus segments based on TRQ measure

Deskriptive Statistik

	N	Spannwei	Minimum	Maximum	Summe	Mittelwert		Standard	Varianz	Schiefe		Kurtosis	
	Statistik	Statistik	Statistik	Statistik	Statistik	Statistik	Standardf ehler	Statistik	Statistik	Statistik	Standardf ehler	Statistik	Standardf ehler
Cw1kCRepQuo	29	,47027	4,82811	5,29838	146,40809	5,0485549	,02081688	,11210232	,013	,003	,434	-,540	,845
Cw1kCcRepQuo	29	1,41513	35,68172	37,09684	1058,140	36,48758	,07613657	,41000798	,168	-,162	,434	-,953	,845
Cw1kCvRepQuo	29	,22020	1,75882	1,97902	53,19269	1,8342308	,01119791	,06030257	,004	,603	,434	-,700	,845
Gültige Werte (Listenweise)	29												

Table 50: Corpus data sets descriptive statistics of $Al1k_{S1}$ corpus segments based on TRQ measure

Deskriptive Statistik

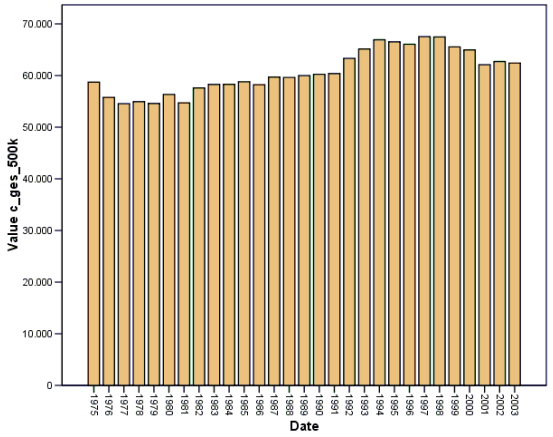
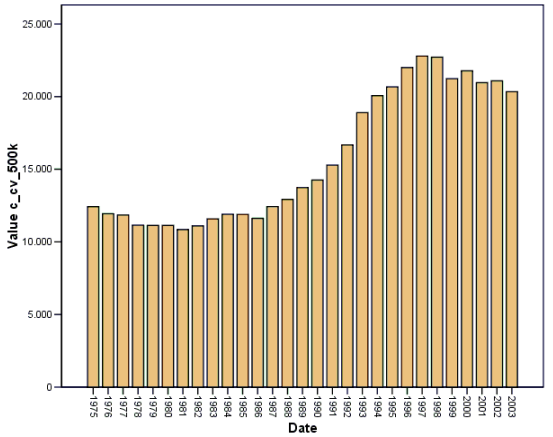
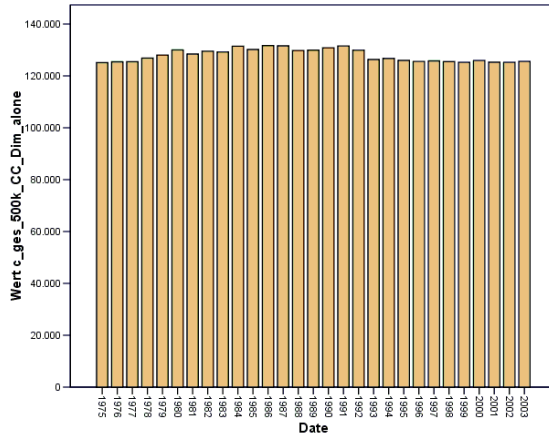
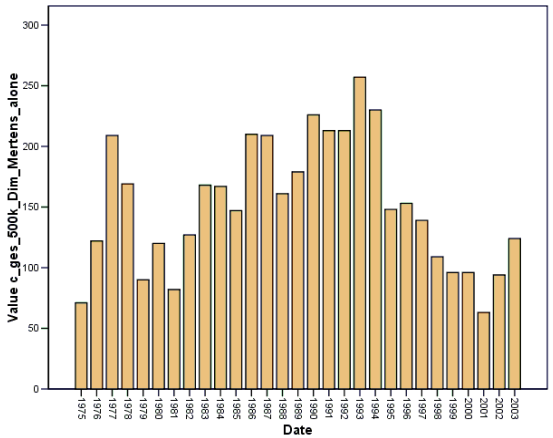
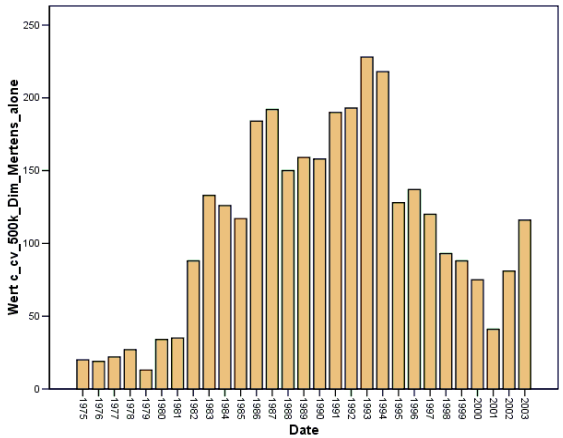
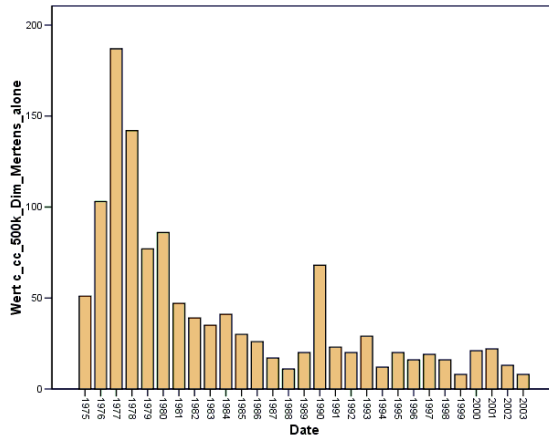
	N	Spannwei	Minimum	Maximum	Summe	Mittelwert		Standard	Varianz	Schiefe		Kurtosis	
	Statistik	Statistik	Statistik	Statistik	Statistik	Statistik	Standardf ehler	Statistik	Statistik	Statistik	Standardf ehler	Statistik	Standardf ehler
Al100S1RepQuo	32	,25152	1,65189	1,90341	56,15295	1,7547797	,01138656	,06441209	,004	,402	,414	-,545	,809
Al100S1CcRepQuo	32	2,72727	10,45455	13,18182	378,54545	11,82955	,12954931	,73284158	,537	,073	,414	-,690	,809
Al100S1CvRepQuo	32	,27268	1,25894	1,53162	43,18022	1,3493820	,00822804	,04654482	,002	1,615	,414	7,087	,809
Gültige Werte (Listenweise)	32												

Table 51: Corpus data sets descriptive statistics of $Al1k_{S2}$ corpus segments based on TRQ measure

Deskriptive Statistik

	N	Spannwei	Minimum	Maximum	Summe	Mittelwert		Standard	Varianz	Schiefe		Kurtosis	
	Statistik	Statistik	Statistik	Statistik	Statistik	Statistik	Standardf ehler	Statistik	Statistik	Statistik	Standardf ehler	Statistik	Standardf ehler
Al100S2CRepQuo	32	,22745	1,64803	1,87547	56,16180	1,7550562	,01090970	,06171456	,004	,433	,414	-,421	,809
Al100S2CcRepQuo	32	3,38095	10,71429	14,09524	393,66667	12,30208	,15148681	,85693882	,734	-,072	,414	-,518	,809
Al100S2CvRepQuo	32	,19210	1,25128	1,44338	43,20847	1,3502648	,00791897	,04479648	,002	,358	,414	,174	,809
Gültige Werte (Listenweise)	32												

Table 52: Overview of number of yearly matched terms by applied taxonomies on CW corpora

Corpus	Taxonomy	C	C_V	C_C
CW_{5k}	Dim			
	Dim_ Mertens_alone			

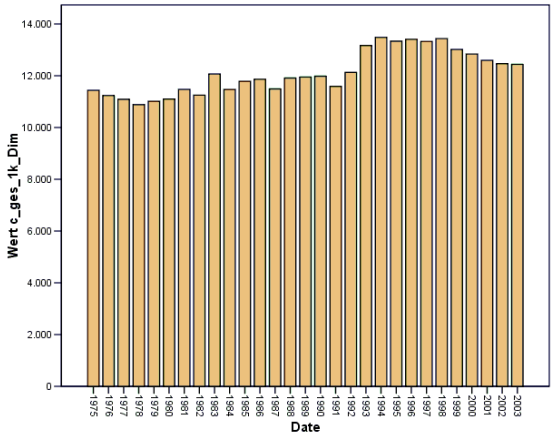
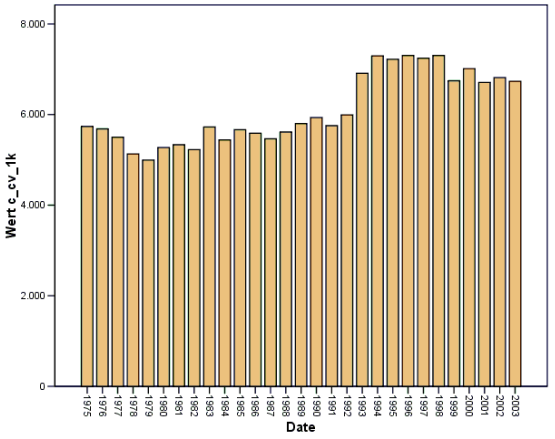
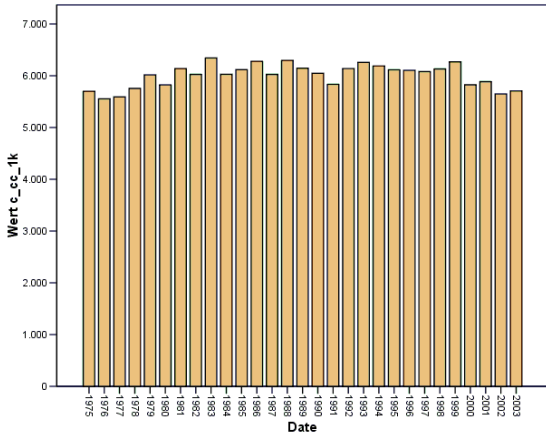
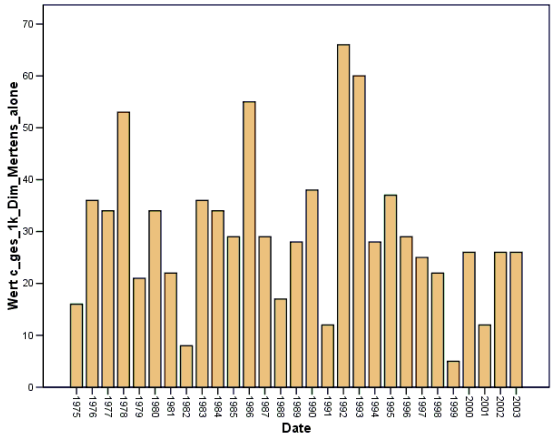
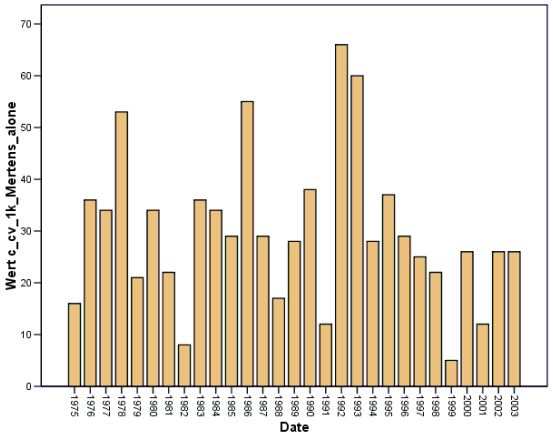
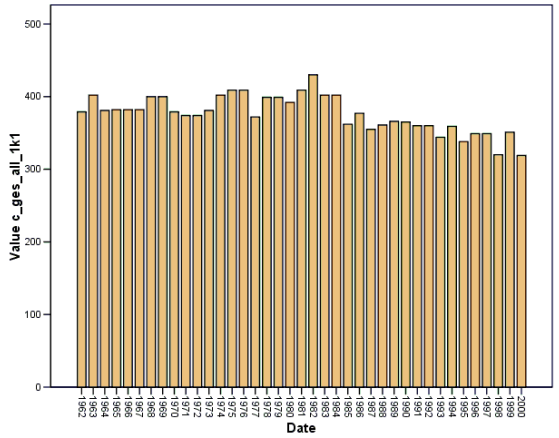
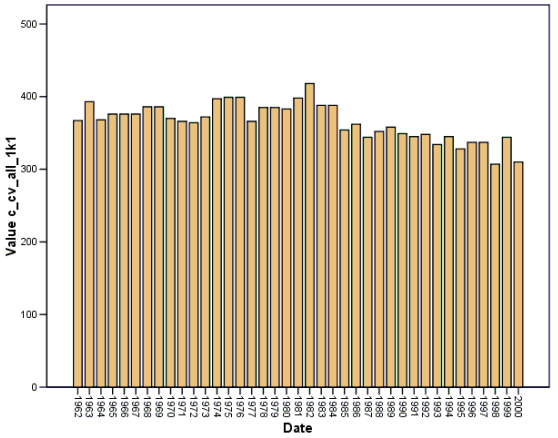
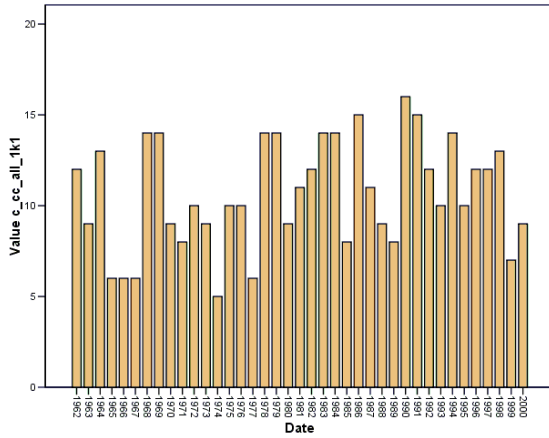
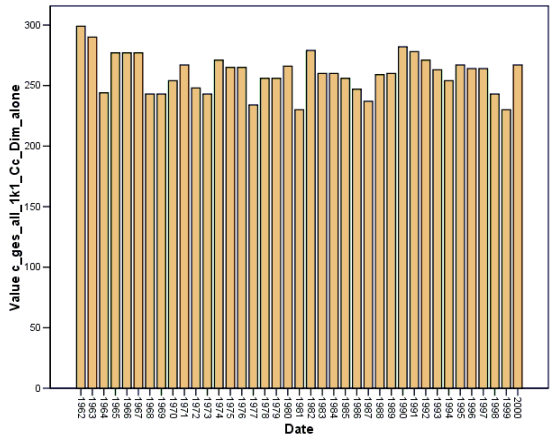
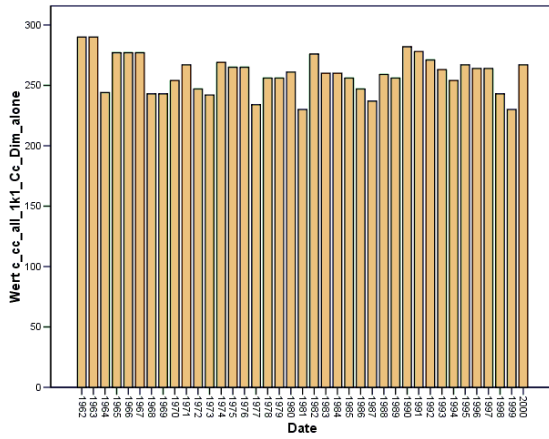
Corpus	Taxonomy	C	C_V	C_C
CW_{1k}	Dim			
	Dim_ Mertens_alone			No terms assigned

Table 53: Overview of number of yearly matched terms by applied taxonomies on Al1k corpora

Corpus	Taxonomy	C	C_V	C_C
$Al1k_{S1}$	Dim			
	CC_Dim_alone		No terms assigned	

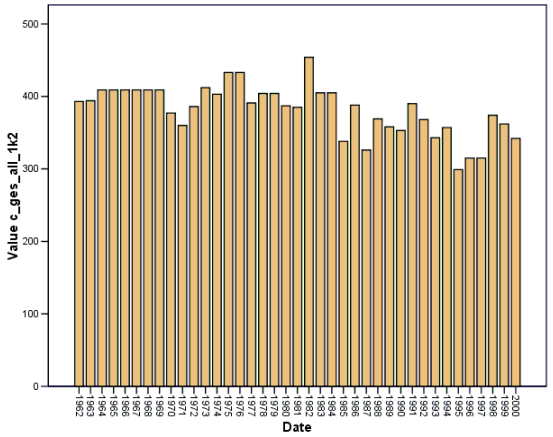
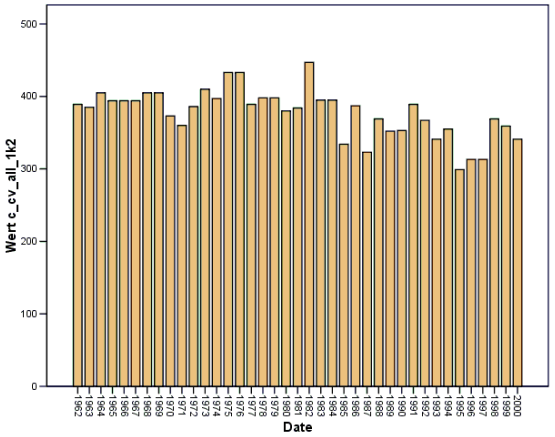
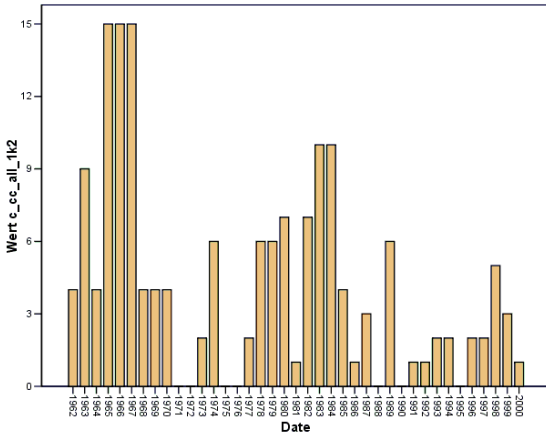
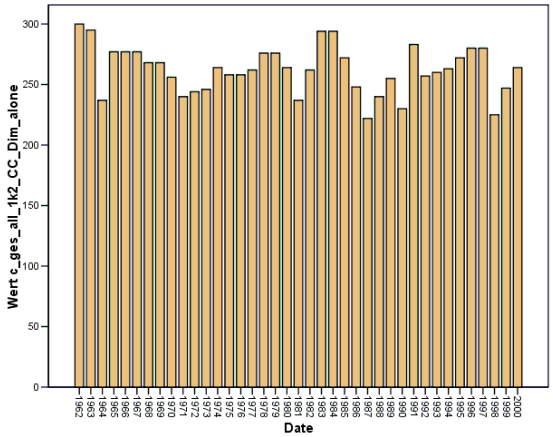
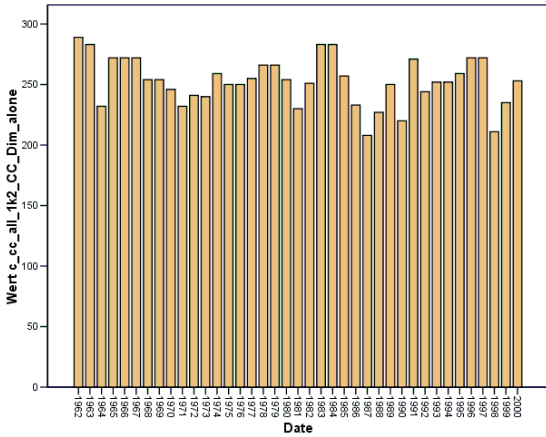
Corpus	Taxonomy	C	C_V	C_C
$Al1k_{S2}$	Dim			
	CC_Dim_alone		No terms assigned	

Table 54: Statistics of terms assigned to dimensions within the three corpus segments C, C_C, C_V

Descriptive Statistics

	N	Range	Minimum	Maximum	Sum	Mean		Std.	Variance	Skewness		Kurtosis	
	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Statistic	Statistic	Std. Error	Statistic	Std. Error
c_ges_500k	29	12982	54541	67523	1761316	60735,03	782,000	4211,201	1,8E+07	,162	,434	-1,178	,845
c_ges_500k_CC_Dim_ alone	29	32176	92941	125117	3243299	111837,90	1542,732	8307,868	6,9E+07	-,780	,434	-,064	,845
c_ges_500k_dim_ mertens	29	201	49	250	3767	129,90	10,040	54,066	2923,096	,503	,434	-,580	,845
c_ges_500k_Dim_ Mertens_alone	29	194	63	257	4392	151,45	9,908	53,357	2846,970	,171	,434	-,988	,845
c_cc_500k	29	6151	41128	47279	1304897	44996,45	324,692	1748,521	3057326	-,661	,434	-,605	,845
c_cc_500k_CC_Dim_ alone	29	32176	92941	125117	3243299	111837,90	1542,732	8307,868	6,9E+07	-,780	,434	-,064	,845
c_cc_500k_Dim_Mertens	29	179	8	187	1207	41,62	7,807	42,040	1767,387	2,164	,434	4,810	,845
c_cc_500k_Dim_ Mertens_alone	29	179	8	187	1207	41,62	7,807	42,040	1767,387	2,164	,434	4,810	,845
c_cv_500k	29	12982	54541	67523	1761316	60735,03	782,000	4211,201	1,8E+07	,162	,434	-1,178	,845
c_cv_500k_Dim_Mertens	29	201	49	250	3767	129,90	10,040	54,066	2923,096	,503	,434	-,580	,845
c_cv_500k_Dim_ Mertens_alone	29	194	63	257	4392	151,45	9,908	53,357	2846,970	,171	,434	-,988	,845
Valid N (listwise)	29												

Table 55: Statistics of terms assigned to dimensions within the three corpus segments C, C_C, C_V

Descriptive Statistics

	N	Range	Minimum	Maximum	Sum	Mean		Std.	Variance	Skewness		Kurtosis	
	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Statistic	Statistic	Std. Error	Statistic	Std. Error
c_ges_1k_Dim	29	2599	10883	13482	351196	12110,21	155,741	838,691	703402,2	,347	,434	-1,210	,845
c_ges_1k_Dim_CC_ Dim_alone	29	6985	18565	25550	654986	22585,72	364,631	1963,600	3855725	-,659	,434	-,711	,845
c_ges_1k_Dim_Mertens	29	55	5	60	766	26,41	2,665	14,354	206,037	,828	,434	,310	,845
c_ges_1k_Dim_ Mertens_alone	29	61	5	66	864	29,79	2,718	14,635	214,170	,755	,434	,642	,845
c_cc_1k	29	790	5553	6343	174053	6001,83	41,664	224,370	50341,719	-,502	,434	-,793	,845
c_cc_1k_CC_Dim_alone	29	6985	18565	25550	654986	22585,72	364,631	1963,600	3855725	-,659	,434	-,711	,845
c_cc_1k_Dim_Mertens	29	0	0	0	0	,00	,000	,000	,000	.	.	.	.
c_cc_1k_Dim_Mertens_ alone	29	0	0	0	0	,00	,000	,000	,000	.	.	.	.
c_cv_1k	29	2309	4995	7304	177143	6108,38	144,093	775,964	602119,8	,393	,434	-1,431	,845
c_cv_1k_Mertens	29	55	5	60	766	26,41	2,665	14,354	206,037	,828	,434	,310	,845
c_cv_1k_Mertens_alone	29	61	5	66	864	29,79	2,718	14,635	214,170	,755	,434	,642	,845
Valid N (listwise)	29												

Descriptive Statistics

	N	Range	Minimum	Maximum	Sum	Mean		Std.	Variance	Skewness		Kurtosis	
	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Statistic	Statistic	Std. Error	Statistic	Std. Error
c_ges_all_1k1	39	111	319	430	14676	376,31	4,055	25,324	641,324	-,280	,378	-,190	,741
c_ges_all_1k1_Cc_Dim	39	11	5	16	416	10,67	,477	2,977	8,860	-,119	,378	-,990	,741
c_ges_all_1k1_Cc_ Dim_alone	39	69	230	299	10146	260,15	2,575	16,078	258,502	,081	,378	-,170	,741
c_cc_all_1k1	39	11	5	16	416	10,67	,477	2,977	8,860	-,119	,378	-,990	,741
c_cc_all_1k1_Cc_Dim	39	11	5	16	416	10,67	,477	2,977	8,860	-,119	,378	-,990	,741
c_cc_all_1k1_Cc_Dim_ alone	39	60	230	290	10121	259,51	2,477	15,466	239,204	-,055	,378	-,496	,741
c_cv_all_1k1	39	111	307	418	14260	365,64	4,056	25,330	641,605	-,329	,378	-,208	,741
Valid N (listwise)	39												

Table 56: Statistics of terms assigned to dimensions within the three corpus segments C, C_C, C_V

Descriptive Statistics

	N	Range	Minimum	Maximum	Sum	Mean		Std.	Variance	Skewness		Kurtosis	
	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Statistic	Statistic	Std. Error	Statistic	Std. Error
c_ges_all_1k2	39	155	299	454	14877	381,46	5,574	34,812	1211,887	-,472	,378	-,059	,741
c_ges_all_1k2_CC_Dim	39	18	3	21	542	13,90	,746	4,661	21,726	-,145	,378	-,861	,741
c_ges_all_1k2_CC_ Dim_alone	39	78	222	300	10228	262,26	3,115	19,451	378,354	-,123	,378	-,443	,741
c_cc_all_1k2	39	15	0	15	164	4,21	,676	4,219	17,799	1,325	,378	1,276	,741
c_cc_all_1k2_Cc_Dim	39	15	0	15	164	4,21	,676	4,219	17,799	1,325	,378	1,276	,741
c_cc_all_1k2_CC_Dim_ alone	39	81	208	289	9850	252,56	3,151	19,679	387,252	-,278	,378	-,247	,741
c_cv_all_1k2	39	148	299	447	14713	377,26	5,344	33,371	1113,617	-,416	,378	,112	,741
Valid N (listwise)	39												

Table 57: Corpus data sets descriptive statistics of CW_{5kb} corpus segments based on TRQ measure

Deskriptive Statistik

	N	Spannwei	Minimum	Maximum	Summe	Mittelwert		Standard	Varianz	Schiefe		Kurtosis	
	Statistik	Statistik	Statistik	Statistik	Statistik	Statistik	Standardf ehler	Statistik	Statistik	Statistik	Standardf ehler	Statistik	Standardf ehler
Cw5kbCRepQuo	29	213,20965	157,77627	370,98592	6060,743	208,9911	10,11433	54,46733	2966,690	1,765	,434	2,684	,845
Cw5kbCcRepQuo	29	2673,203	1667,932	4341,135	67309,21	2321,007	123,2577	663,7633	440581,7	1,850	,434	2,982	,845
Cw5kbCvRepQuo	29	21,17990	17,04783	38,22773	650,41276	22,42803	1,148753	6,186222	38,269	1,500	,434	1,049	,845
Gültige Werte (Listenweise)	29												

Table 58: Corpus data sets descriptive statistics of CW_{5kbu} corpus segments based on TRQ measure

Deskriptive Statistik

	N	Spannwei	Minimum	Maximum	Summe	Mittelwert		Standard	Varianz	Schiefe		Kurtosis	
	Statistik	Statistik	Statistik	Statistik	Statistik	Statistik	Standardf ehler	Statistik	Statistik	Statistik	Standardf ehler	Statistik	Standardf ehler
Cw5kbuCRepQuo	29	203,87522	153,99943	357,87464	5903,803	203,5794	9,738231	52,44198	2750,161	1,679	,434	2,433	,845
Cw5kbuCcRepQuo	29	2288,251	1427,593	3715,843	57638,36	1987,530	105,5392	568,3459	323017,1	1,843	,434	2,960	,845
Cw5kbuCvRepQuo	29	20,13000	15,90145	36,03145	617,78280	21,30286	1,092966	5,885804	34,643	1,397	,434	,814	,845
Gültige Werte (Listenweise)	29												

Table 59: Corpus data sets descriptive statistics of CW_{5kbun} corpus segments based on TRQ measure

Deskriptive Statistik

	N	Spannwei	Minimum	Maximum	Summe	Mittelwert		Standard	Varianz	Schiefe		Kurtosis	
	Statistik	Statistik	Statistik	Statistik	Statistik	Statistik	Standardf ehler	Statistik	Statistik	Statistik	Standardf ehler	Statistik	Standardf ehler
Cw5kbunCRepQuo	29	56,20468	49,79404	105,99873	1837,336	63,35642	2,329577	12,54516	157,381	2,131	,434	4,851	,845
Cw5kbunCcRepQuo	29	30,23133	516,53978	546,77111	15647,08	539,5546	,91245840	4,913739	24,145	-3,776	,434	18,212	,845
Cw5kbunCvRepQuo	29	7,08353	6,57903	13,66256	243,07962	8,3820558	,34258570	1,844880	3,404	1,835	,434	2,517	,845
Gültige Werte (Listenweise)	29												

Table 60: Corpus data sets descriptive statistics of CW_{5k} corpus segments based on TRQ measure

Deskriptive Statistik

	N	Spannwei	Minimum	Maximum	Summe	Mittelwert		Standard	Varianz	Schiefe		Kurtosis	
	Statistik	Statistik	Statistik	Statistik	Statistik	Statistik	Standardf ehler	Statistik	Statistik	Statistik	Standardf ehler	Statistik	Standardf ehler
Cw5kbun2CRepQuo	29	40,01640	51,98305	91,99945	1860,828	64,16648	1,898940	10,22611	104,573	1,110	,434	1,039	,845
Cw5kbun2CcRepQuo	29	24,23105	510,79061	535,02166	15411,26	531,4226	,80747527	4,348387	18,908	-4,013	,434	19,184	,845
Cw5kbun2CvRepQuo	29	5,36104	6,82736	12,18839	243,85316	8,4087296	,29450024	1,585932	2,515	1,288	,434	,819	,845
Gültige Werte (Listenweise)	29												

Table 61: Corpus data sets descriptive statistics of CW_{1kb} corpus segments based on TRQ measure

Deskriptive Statistik

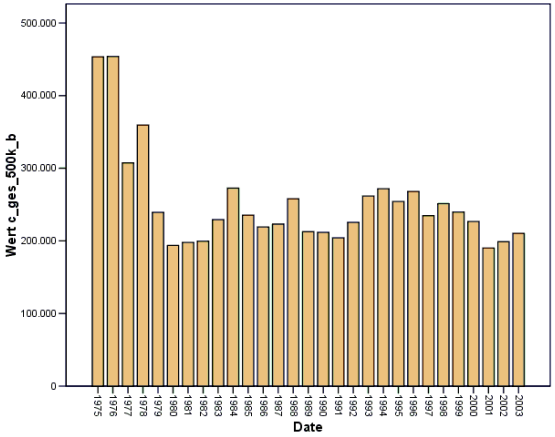
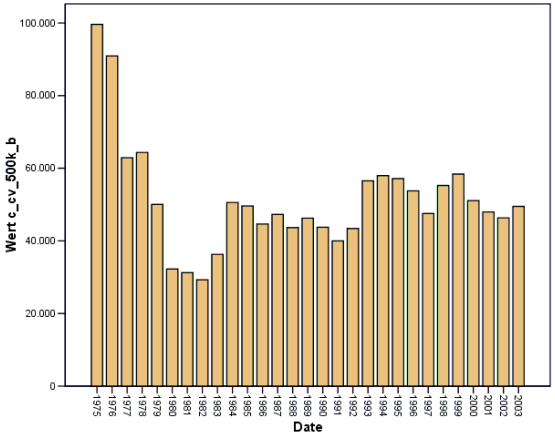
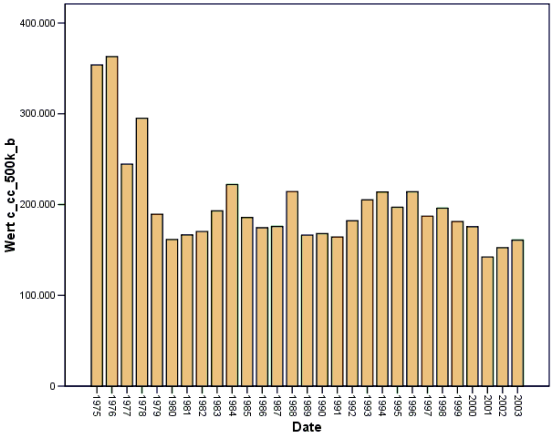
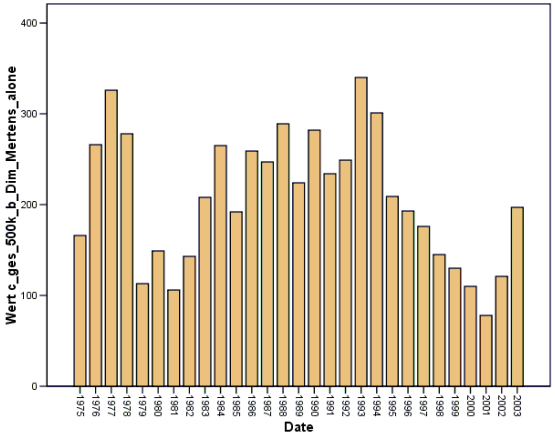
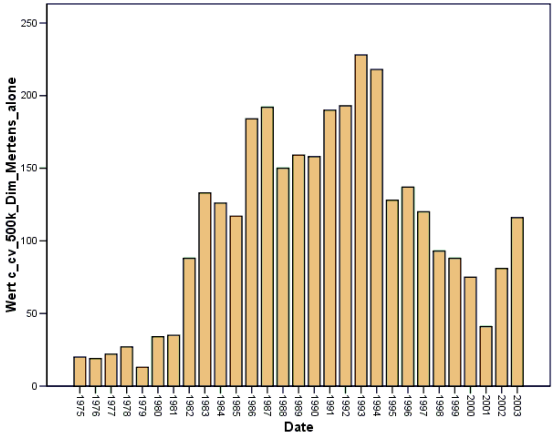
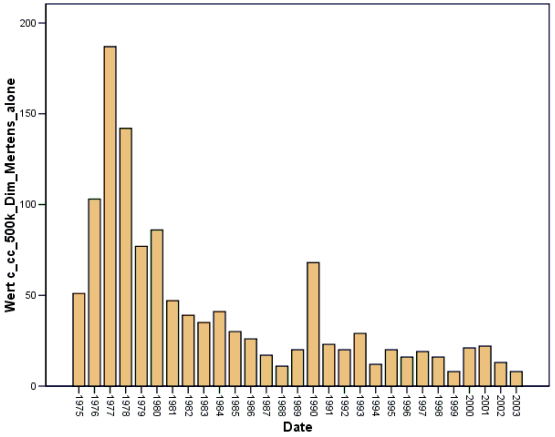
	N	Spannwei	Minimum	Maximum	Summe	Mittelwert		Standard	Varianz	Schiefe		Kurtosis	
	Statistik	Statistik	Statistik	Statistik	Statistik	Statistik	Standardf ehler	Statistik	Statistik	Statistik	Standardf ehler	Statistik	Standardf ehler
Cw1kbCRepQuo	29	140,15545	92,00150	232,15695	3586,570	123,6748	6,603390	35,56034	1264,538	1,813	,434	3,022	,845
Cw1kbCcRepQuo	29	1645,642	916,25452	2561,897	37188,89	1282,376	75,71987	407,7640	166271,5	1,898	,434	3,404	,845
Cw1kbCvRepQuo	29	15,38388	11,10707	26,49095	429,38782	14,80648	,78666437	4,236317	17,946	1,592	,434	1,600	,845
Gültige Werte (Listenweise)	29												

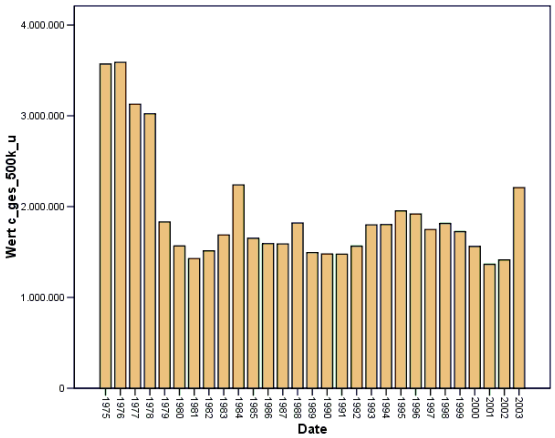
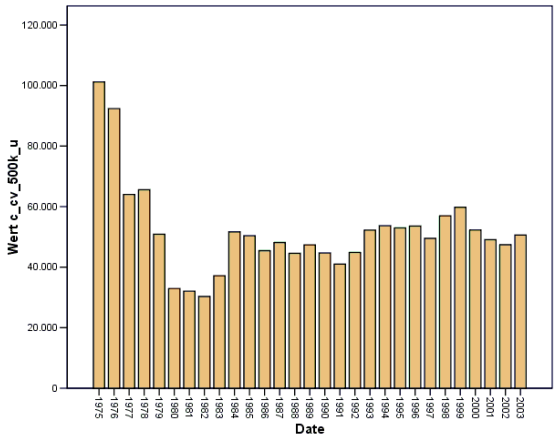
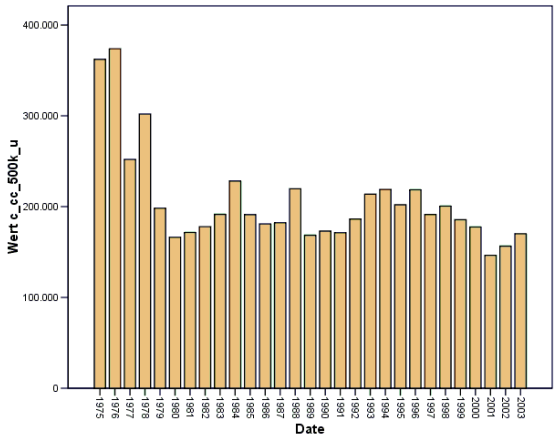
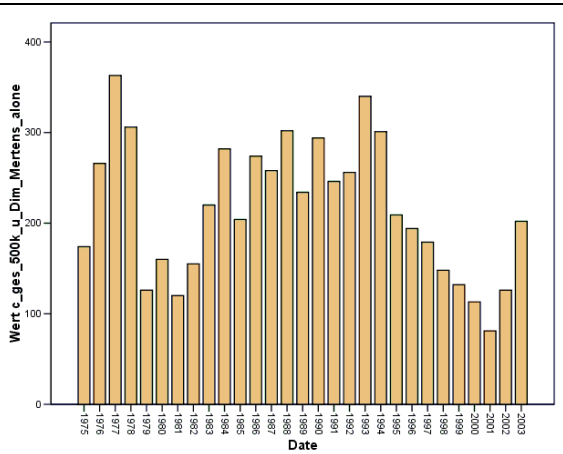
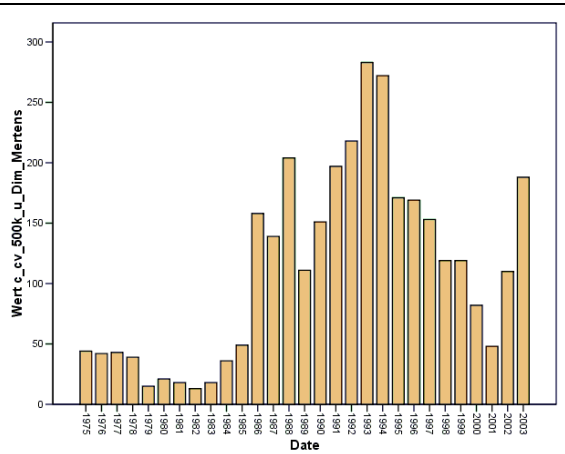
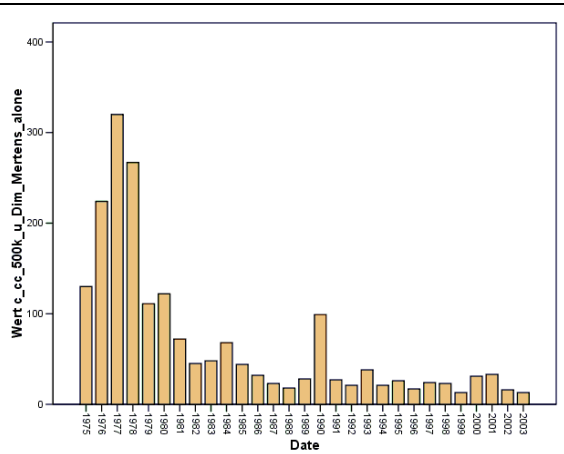
Table 62: Corpus data sets descriptive statistics of CW_{1kbu} corpus segments based on TRQ measure

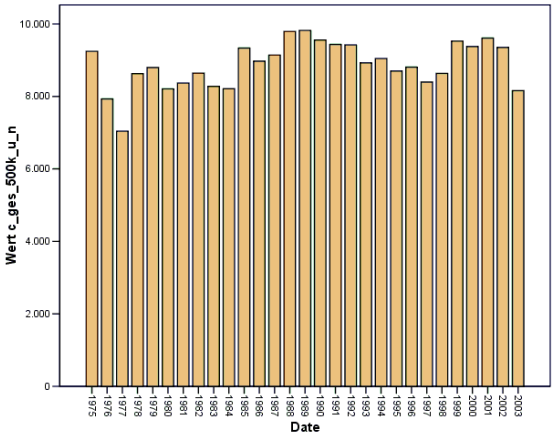
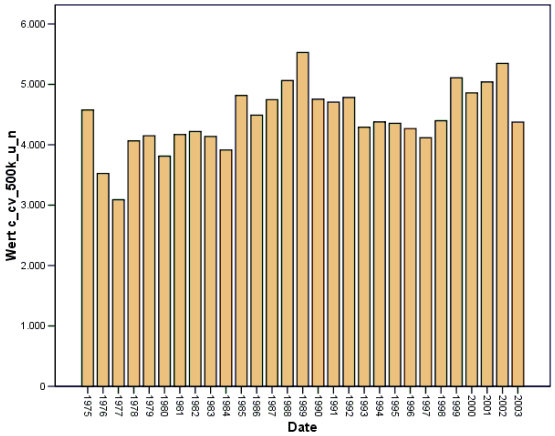
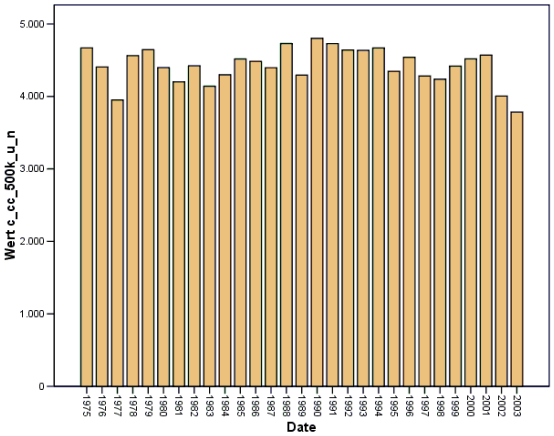
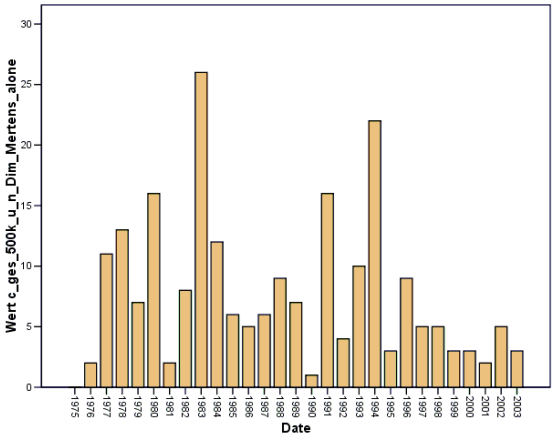
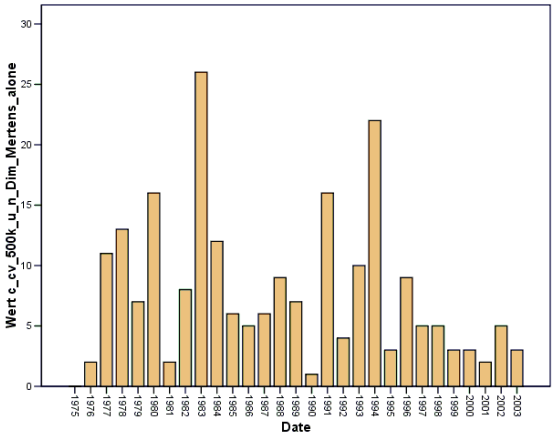
Deskriptive Statistik

	N	Spannwei	Minimum	Maximum	Summe	Mittelwert		Standard	Varianz	Schiefe		Kurtosis	
	Statistik	Statistik	Statistik	Statistik	Statistik	Statistik	Standardf ehler	Statistik	Statistik	Statistik	Standardf ehler	Statistik	Standardf ehler
Cw1kbuCRepQuo	29	135,90297	90,11391	226,01687	3517,894	121,3067	6,439528	34,67792	1202,558	1,763	,434	2,852	,845
Cw1kbuCcRepQuo	29	1473,490	821,11211	2294,602	33321,99	1149,034	67,81138	365,1755	133353,1	1,896	,434	3,398	,845
Cw1kbuCvRepQuo	29	14,63698	10,37334	25,01031	407,07281	14,03699	,75175306	4,048314	16,389	1,486	,434	1,317	,845
Gültige Werte (Listenweise)	29												

Table 63: Overview of number of yearly matched terms by applied taxonomies on CW corpora

Corpus	Taxon-omy	C	C_V	C_C
CW_{5kb}	Dim			
	Dim_ Mertens _alone			

Corpus	Taxon-omy	C	C_V	C_C
CW_{5kbu}	Dim			
	Dim_ Mertens _alone			

Corpus	Taxon-omy	C	C_V	C_C
CW_{5kbun}	Dim			
	Dim_ Mertens _alone			No terms assigned

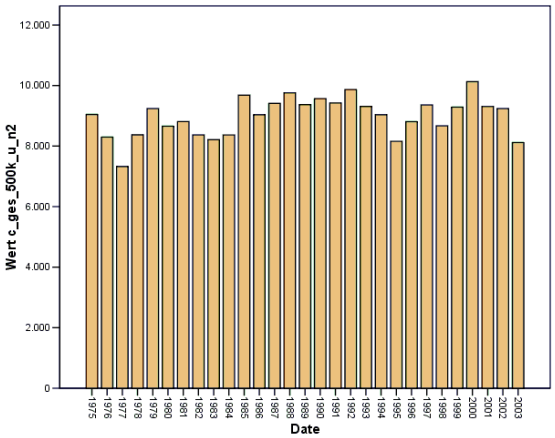
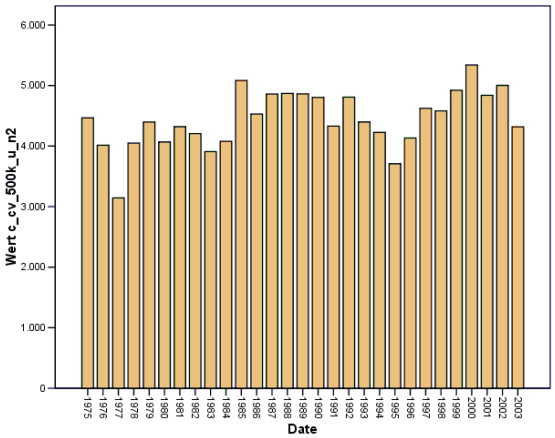
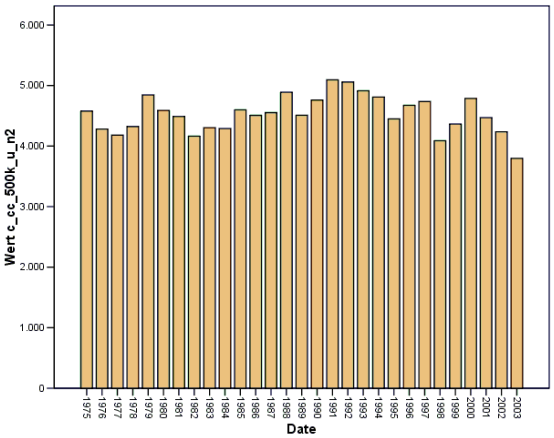
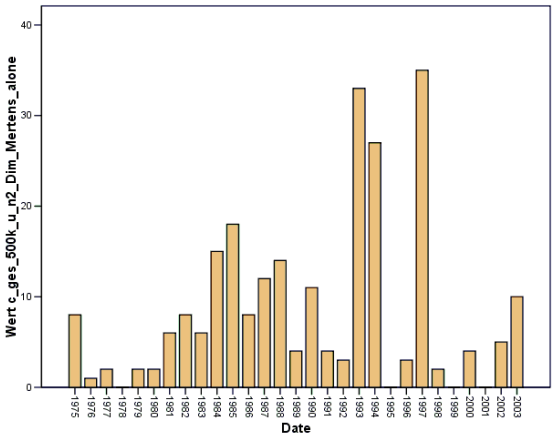
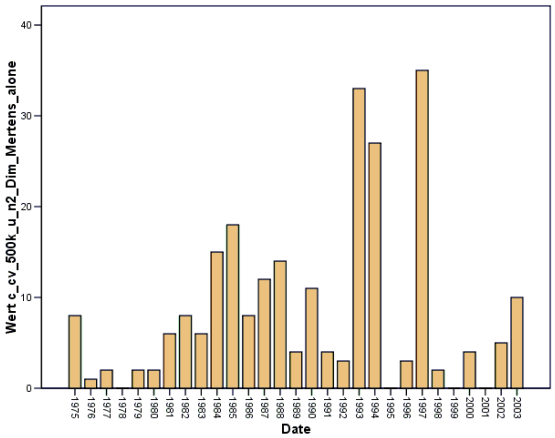
Corpus	Taxon-omy	C	C_V	C_C
CW_{5kbun2}	Dim			
	Dim_ Mertens _alone			No terms assigned

Table 64: Statistics of terms assigned to dimensions within the three corpus segments C, C_C, C_V

Descriptive Statistics

	N	Range	Minimum	Maximum	Sum	Mean		Std.	Variance	Skewness		Kurtosis	
	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Statistic	Statistic	Std. Error	Statistic	Std. Error
c_ges_500k_b	29	263762	190155	453917	7303245	251836,03	12393,813	66742,725	4,5E+09	2,139	,434	4,518	,845
c_ges_500k_b_CC_ Dim_alone	29	228748	154846	383594	6144889	211892,72	9330,376	50245,612	2,5E+09	2,355	,434	5,778	,845
c_ges_500k_b_ Dim_Mertens	29	275	51	326	5126	176,76	14,420	77,655	6030,261	,330	,434	-,760	,845
c_ges_500k_b_ Dim_Mertens_alone	29	262	78	340	5996	206,76	13,358	71,935	5174,618	,020	,434	-1,011	,845
c_cc_500k_b	29	220797	142199	362996	5815815	200545,34	9891,212	53265,808	2,8E+09	2,067	,434	4,134	,845
c_cc_500k_b_CC_ Dim_alone	29	228748	154846	383594	6127017	211276,45	9339,827	50296,506	2,5E+09	2,386	,434	5,875	,845
c_cc_500k_b_Dim_ Mertens	29	272	12	284	1784	61,52	13,321	71,735	5145,973	2,092	,434	3,684	,845
c_cc_500k_b_Dim_ Mertens_alone	29	272	12	284	1784	61,52	13,321	71,735	5145,973	2,092	,434	3,684	,845
c_cv_500k_b	29	70347	29266	99613	1487430	51290,69	2788,463	15016,331	2,3E+08	1,670	,434	4,121	,845
c_cv_500k_b_Dim_ Mertens	29	275	13	288	3342	115,24	15,122	81,433	6631,261	,451	,434	-,805	,845
c_cv_500k_b_Dim_ Mertens_alone	29	287	15	302	4212	145,24	15,541	83,690	7004,047	,066	,434	-1,109	,845
Valid N (listwise)	29												

Table 65: Statistics of terms assigned to dimensions within the three corpus segments C, C_C, C_V

Descriptive Statistics

	N	Range	Minimum	Maximum	Sum	Mean		Std.	Variance	Skewness		Kurtosis	
	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Statistic	Statistic	Std. Error	Statistic	Std. Error
c_ges_500k_u	29	270677	195455	466132	7480132	257935,59	12625,161	67988,572	4,6E+09	2,187	,434	4,692	,845
c_ges_500k_u_CC_Dim_alone	29	237543	165449	402992	6403947	220825,76	9946,060	53561,172	2,9E+09	2,424	,434	5,953	,845
c_ges_500k_u_Dim_Mertens	29	0	0	0	0	,00	,000	,000	,000	.	.	.	.
c_ges_500k_u_Dim_Mertens_alone	29	282	81	363	6265	216,03	13,923	74,978	5621,677	,087	,434	-,946	,845
c_cc_500k_u	29	227367	146393	373760	5977511	206121,07	10112,779	54458,980	3,0E+09	2,087	,434	4,218	,845
c_cc_500k_u_CC_Dim_alone	29	237543	165449	402992	6403947	220825,76	9946,060	53561,172	2,9E+09	2,424	,434	5,953	,845
c_cc_500k_u_Dim_Mertens	29	307	13	320	1954	67,38	14,580	78,516	6164,815	2,131	,434	4,060	,845
c_cc_500k_u_Dim_Mertens_alone	29	307	13	320	1954	67,38	14,580	78,516	6164,815	2,131	,434	4,060	,845
c_cv_500k_u	29	70855	30330	101185	1502621	51814,52	2790,263	15026,026	2,3E+08	1,829	,434	4,701	,845
c_cv_500k_u_Dim_Mertens	29	270	13	283	3230	111,38	14,886	80,163	6426,030	,480	,434	-,762	,845
c_cv_500k_u_Dim_Mertens_alone	29	287	15	302	4311	148,66	15,945	85,866	7373,020	,025	,434	-1,189	,845
Valid N (listwise)	29												

Correlations

		LanInd	All100S1_ Article	All100S1_ Conjunction	All100S1_ Particle	All100S1_ Preposition	All100S1_ Pronom	All100S1_ Verb
LanInd	Pearson Correlation	1	,764**	-,452**	,339	-,729**	,631**	,073
	Sig. (2-tailed)		,000	,009	,058	,000	,000	,693
	N	32	32	32	32	32	32	32
All100S1_Article	Pearson Correlation	,764**	1	-,617**	,433*	-,540**	,429*	,037
	Sig. (2-tailed)	,000		,000	,013	,001	,014	,842
	N	32	32	32	32	32	32	32
All100S1_Conjunction	Pearson Correlation	-,452**	-,617**	1	-,155	,253	-,130	,147
	Sig. (2-tailed)	,009	,000		,398	,163	,479	,424
	N	32	32	32	32	32	32	32
All100S1_Particle	Pearson Correlation	,339	,433*	-,155	1	-,306	-,001	,342
	Sig. (2-tailed)	,058	,013	,398		,088	,998	,056
	N	32	32	32	32	32	32	32
All100S1_Preposition	Pearson Correlation	-,729**	-,540**	,253	-,306	1	-,399*	-,069
	Sig. (2-tailed)	,000	,001	,163	,088		,024	,709
	N	32	32	32	32	32	32	32
All100S1_Pronom	Pearson Correlation	,631**	,429*	-,130	-,001	-,399*	1	,008
	Sig. (2-tailed)	,000	,014	,479	,998	,024		,965
	N	32	32	32	32	32	32	32
All100S1_Verb	Pearson Correlation	,073	,037	,147	,342	-,069	,008	1
	Sig. (2-tailed)	,693	,842	,424	,056	,709	,965	
	N	32	32	32	32	32	32	32

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

Fig. 114: Correlations between LanInd and CountSum measure in $Al1k_{S1}$ corpus test set for each single concept that constitutes Dimension CC_Dim

Correlations

		LanInd	All100S2_ Article	All100S2_ Conjunction	All100S2_ Particle	All100S2_ Preposition	All100S2_ Pronom	All100S2_ Verb
LanInd	Pearson Correlation	1	,821**	,092	,269	-,733**	,557**	-,244
	Sig. (2-tailed)		,000	,618	,137	,000	,001	,178
	N	32	32	32	32	32	32	32
All100S2_Article	Pearson Correlation	,821**	1	-,164	,190	-,640**	,492**	-,160
	Sig. (2-tailed)	,000		,370	,299	,000	,004	,381
	N	32	32	32	32	32	32	32
All100S2_Conjunction	Pearson Correlation	,092	-,164	1	,080	-,022	-,086	-,046
	Sig. (2-tailed)	,618	,370		,665	,906	,641	,804
	N	32	32	32	32	32	32	32
All100S2_Particle	Pearson Correlation	,269	,190	,080	1	-,118	,152	,096
	Sig. (2-tailed)	,137	,299	,665		,522	,408	,602
	N	32	32	32	32	32	32	32
All100S2_Preposition	Pearson Correlation	-,733**	-,640**	-,022	-,118	1	-,417*	,242
	Sig. (2-tailed)	,000	,000	,906	,522		,017	,182
	N	32	32	32	32	32	32	32
All100S2_Pronom	Pearson Correlation	,557**	,492**	-,086	,152	-,417*	1	-,166
	Sig. (2-tailed)	,001	,004	,641	,408	,017		,363
	N	32	32	32	32	32	32	32
All100S2_Verb	Pearson Correlation	-,244	-,160	-,046	,096	,242	-,166	1
	Sig. (2-tailed)	,178	,381	,804	,602	,182	,363	
	N	32	32	32	32	32	32	32

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

Fig. 115: Correlations between LanInd and CountSum measure in $Al1k_{S2}$ corpus test set for each single concept that constitutes Dimension CC_Dim

Table 66: Correlations between CountSum measures in $Al1k_{S1}$ and $Al1k_{S2}$ German source language corpus test sets for each single concept that constitutes Dimension CC_Dim

Korrelationen

		All100S1_ Article	All100S1_ Conjunction	All100S1_ Particle	All100S1_ Preposition	All100S1_ Pronom	All100S1_ Verb
All100S1_Article	Korrelation nach Pearson	1	-,493*	,367	,175	-,143	,042
	Signifikanz (2-seitig)		,014	,078	,413	,505	,847
	N	24	24	24	24	24	24
All100S1_Conjunction	Korrelation nach Pearson	-,493*	1	-,020	-,451*	,153	,268
	Signifikanz (2-seitig)	,014		,927	,027	,477	,205
	N	24	24	24	24	24	24
All100S1_Particle	Korrelation nach Pearson	,367	-,020	1	-,137	-,306	,390
	Signifikanz (2-seitig)	,078	,927		,522	,146	,059
	N	24	24	24	24	24	24
All100S1_Preposition	Korrelation nach Pearson	,175	-,451*	-,137	1	,072	-,193
	Signifikanz (2-seitig)	,413	,027	,522		,739	,367
	N	24	24	24	24	24	24
All100S1_Pronom	Korrelation nach Pearson	-,143	,153	-,306	,072	1	-,071
	Signifikanz (2-seitig)	,505	,477	,146	,739		,741
	N	24	24	24	24	24	24
All100S1_Verb	Korrelation nach Pearson	,042	,268	,390	-,193	-,071	1
	Signifikanz (2-seitig)	,847	,205	,059	,367	,741	
	N	24	24	24	24	24	24

*. Die Korrelation ist auf dem Niveau von 0,05 (2-seitig) signifikant.

Korrelationen

		All100S2_ Article	All100S2_ Conjunction	All100S2_ Particle	All100S2_ Preposition	All100S2_ Pronom	All100S2_ Verb
All100S2_Article	Korrelation nach Pearson	1	-,429*	-,070	-,024	,082	,153
	Signifikanz (2-seitig)		,037	,746	,910	,704	,476
	N	24	24	24	24	24	24
All100S2_Conjunction	Korrelation nach Pearson	-,429*	1	,032	-,082	-,132	-,026
	Signifikanz (2-seitig)	,037		,884	,703	,540	,904
	N	24	24	24	24	24	24
All100S2_Particle	Korrelation nach Pearson	-,070	,032	1	,125	,016	,179
	Signifikanz (2-seitig)	,746	,884		,562	,941	,402
	N	24	24	24	24	24	24
All100S2_Preposition	Korrelation nach Pearson	-,024	-,082	,125	1	,148	-,272
	Signifikanz (2-seitig)	,910	,703	,562		,490	,198
	N	24	24	24	24	24	24
All100S2_Pronom	Korrelation nach Pearson	,082	-,132	,016	,148	1	,059
	Signifikanz (2-seitig)	,704	,540	,941	,490		,783
	N	24	24	24	24	24	24
All100S2_Verb	Korrelation nach Pearson	,153	-,026	,179	-,272	,059	1
	Signifikanz (2-seitig)	,476	,904	,402	,198	,783	
	N	24	24	24	24	24	24

*. Die Korrelation ist auf dem Niveau von 0,05 (2-seitig) signifikant.

Table 67: Correlations between CountSum measures in $AI1k_{S1}$ and $AI1k_{S2}$ English source language corpus test set for each single concept that constitutes Dimension CC_Dim

Korrelationen

		All100S1_ Article	All100S1_ Conjunction	All100S1_ Particle	All100S1_ Preposition	All100S1_ Pronom	All100S1_ Verb
All100S1_Article	Korrelation nach Pearson	1	-,457	-,103	-,389	,081	-,306
	Signifikanz (2-seitig)		,255	,808	,341	,848	,461
	N	8	8	8	8	8	8
All100S1_Conjunction	Korrelation nach Pearson	-,457	1	,074	,543	,623	,075
	Signifikanz (2-seitig)	,255		,863	,164	,099	,859
	N	8	8	8	8	8	8
All100S1_Particle	Korrelation nach Pearson	-,103	,074	1	,227	-,129	-,118
	Signifikanz (2-seitig)	,808	,863		,588	,761	,781
	N	8	8	8	8	8	8
All100S1_Preposition	Korrelation nach Pearson	-,389	,543	,227	1	,409	,775*
	Signifikanz (2-seitig)	,341	,164	,588		,314	,024
	N	8	8	8	8	8	8
All100S1_Pronom	Korrelation nach Pearson	,081	,623	-,129	,409	1	,138
	Signifikanz (2-seitig)	,848	,099	,761	,314		,744
	N	8	8	8	8	8	8
All100S1_Verb	Korrelation nach Pearson	-,306	,075	-,118	,775*	,138	1
	Signifikanz (2-seitig)	,461	,859	,781	,024	,744	
	N	8	8	8	8	8	8

*. Die Korrelation ist auf dem Niveau von 0,05 (2-seitig) signifikant.

Korrelationen

		All100S2_ Article	All100S2_ Conjunction	All100S2_ Particle	All100S2_ Preposition	All100S2_ Pronom	All100S2_ Verb
All100S2_Article	Korrelation nach Pearson	1	-,422	,146	-,430	,014	-,256
	Signifikanz (2-seitig)		,298	,731	,288	,973	,541
	N	8	8	8	8	8	8
All100S2_Conjunction	Korrelation nach Pearson	-,422	1	,322	,442	-,323	-,019
	Signifikanz (2-seitig)	,298		,437	,273	,436	,964
	N	8	8	8	8	8	8
All100S2_Particle	Korrelation nach Pearson	,146	,322	1	,196	-,162	,260
	Signifikanz (2-seitig)	,731	,437		,641	,701	,534
	N	8	8	8	8	8	8
All100S2_Preposition	Korrelation nach Pearson	-,430	,442	,196	1	-,575	,815*
	Signifikanz (2-seitig)	,288	,273	,641		,136	,014
	N	8	8	8	8	8	8
All100S2_Pronom	Korrelation nach Pearson	,014	-,323	-,162	-,575	1	-,382
	Signifikanz (2-seitig)	,973	,436	,701	,136		,350
	N	8	8	8	8	8	8
All100S2_Verb	Korrelation nach Pearson	-,256	-,019	,260	,815*	-,382	1
	Signifikanz (2-seitig)	,541	,964	,534	,014	,350	
	N	8	8	8	8	8	8

*. Die Korrelation ist auf dem Niveau von 0,05 (2-seitig) signifikant.

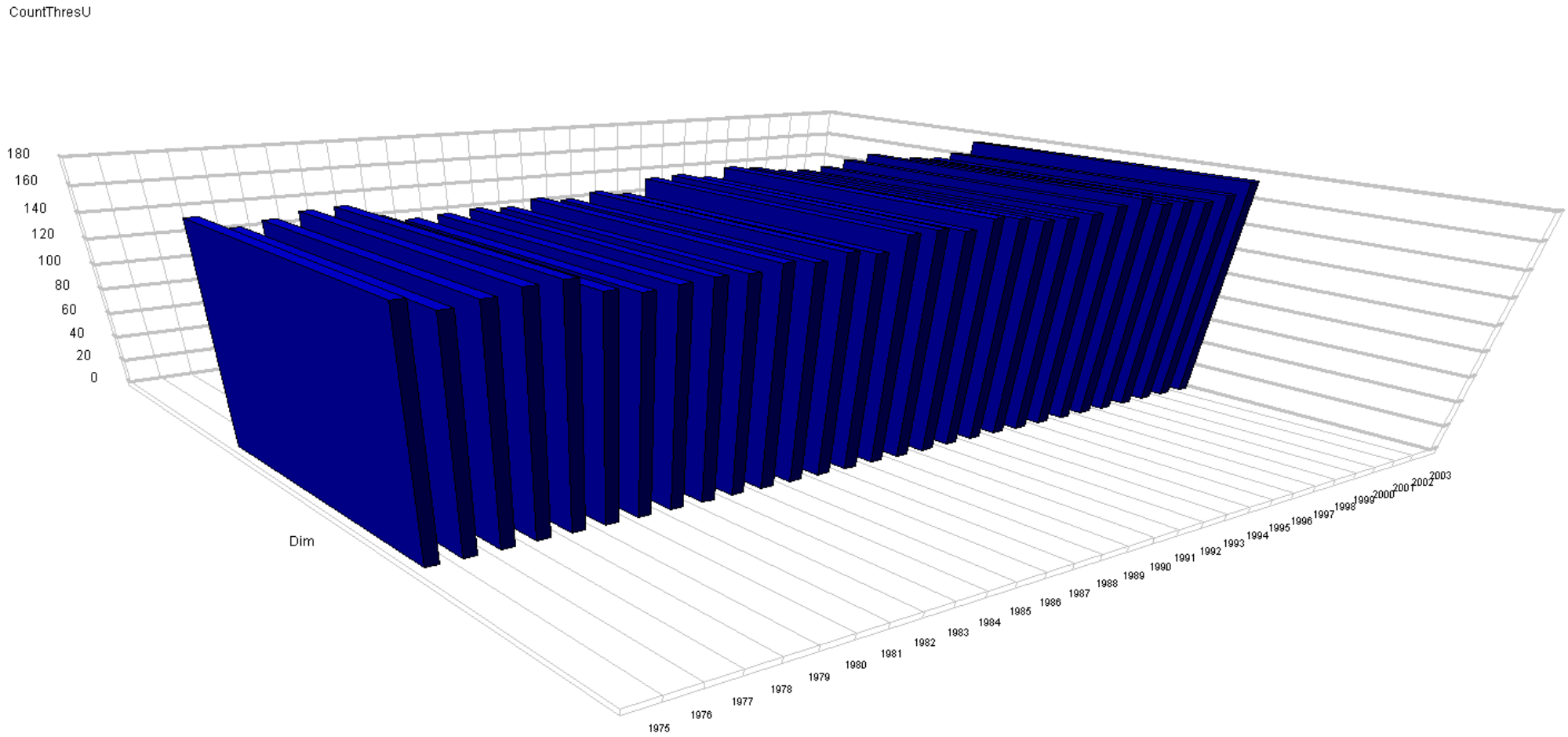


Fig. 116: Visualization of the start node of “Dim” taxonomy for each yearly segment with TRQ threshold filtering for corpus segment C_C of test set CW_{5k}

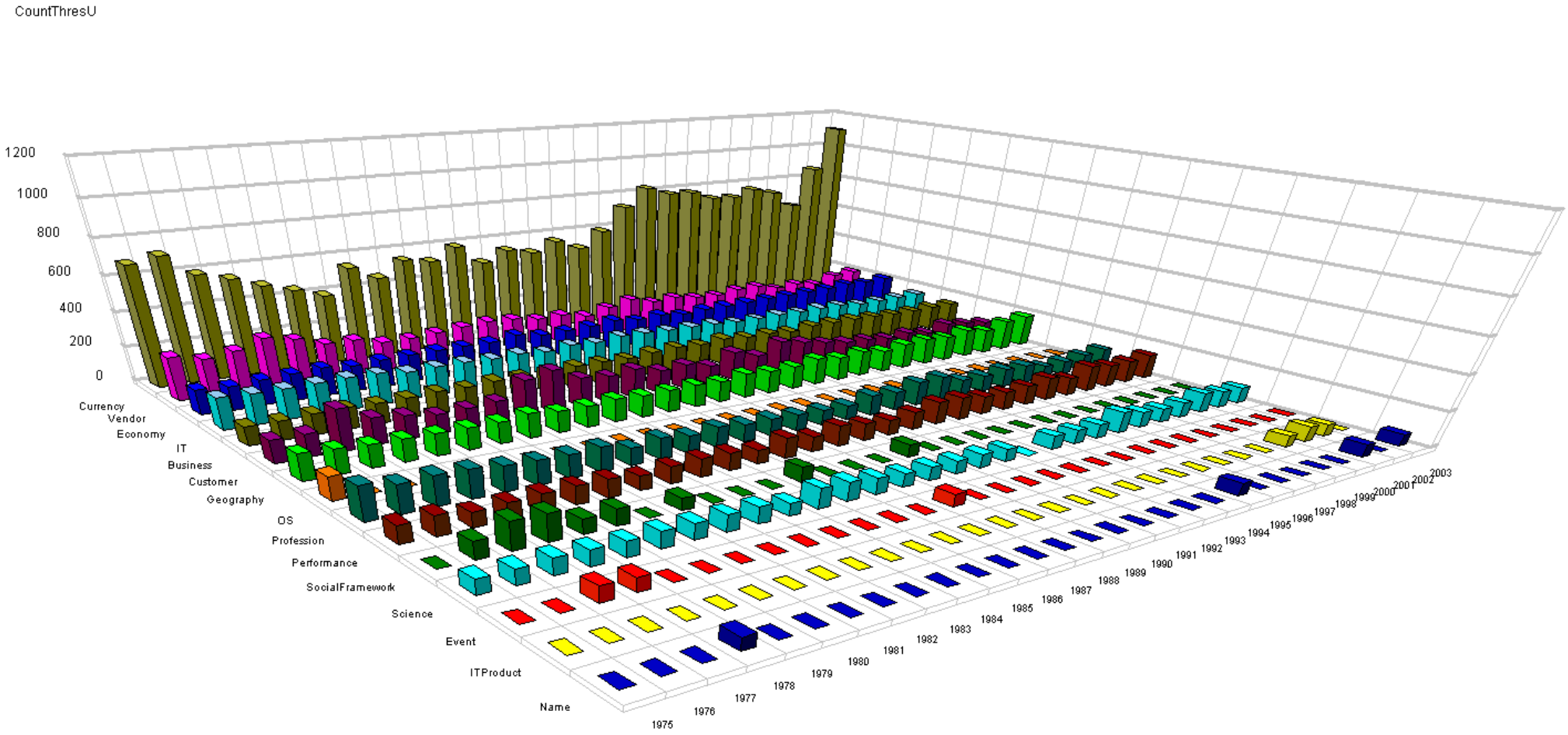


Fig. 117: Visualization of the 2nd level of "Dim" taxonomy for each yearly segment with TRQ threshold filtering for corpus segment C_C of test set CW_{5k}

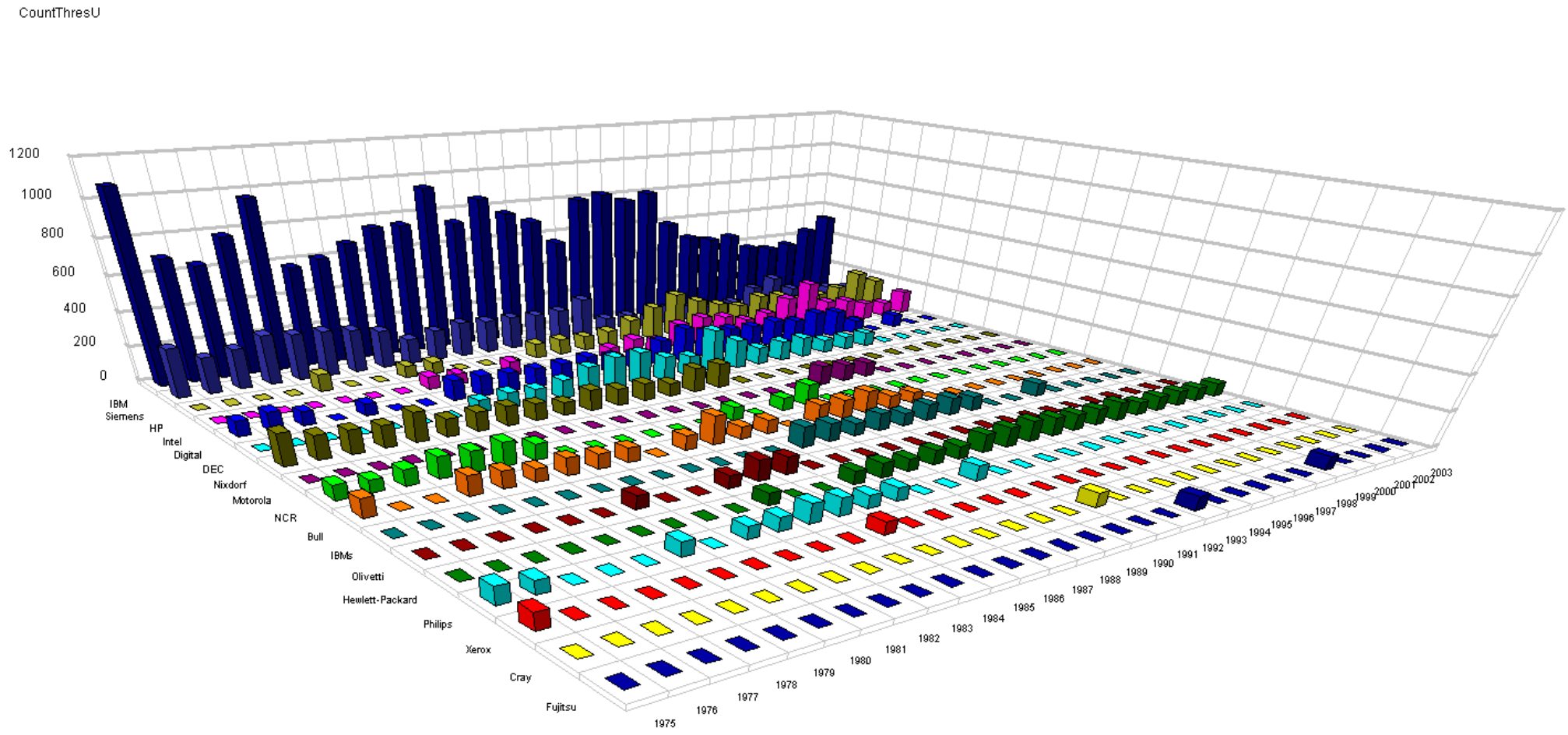


Fig. 118: Visualization of the drill-down within dimension “Vendor” for each yearly segment with TRQ threshold filtering for corpus segment C_C of test set CW_{5k}

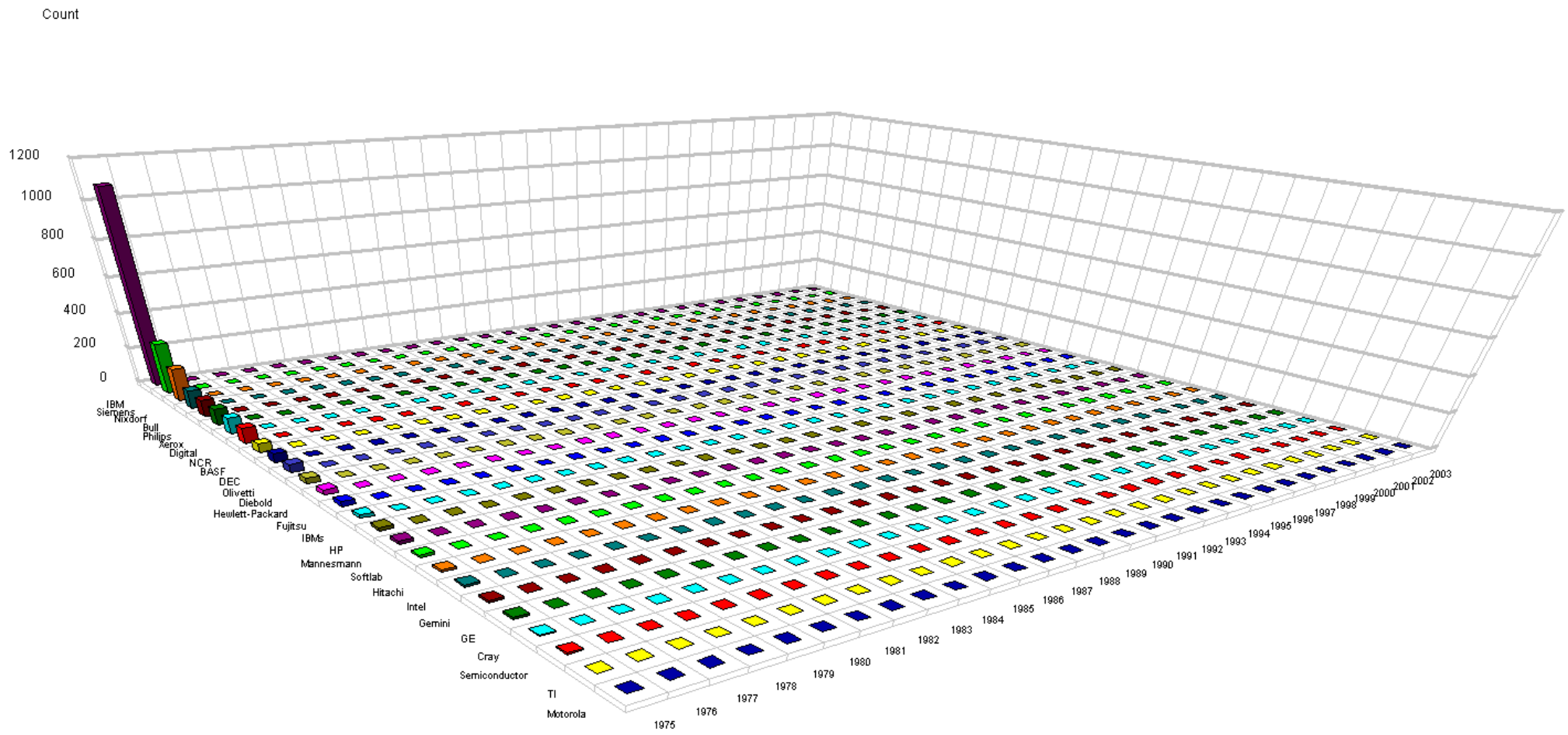


Fig. 119: Visualization of the drill-down dimension “Vendor” for segment “1975” with simple Count measure (without threshold filtering)

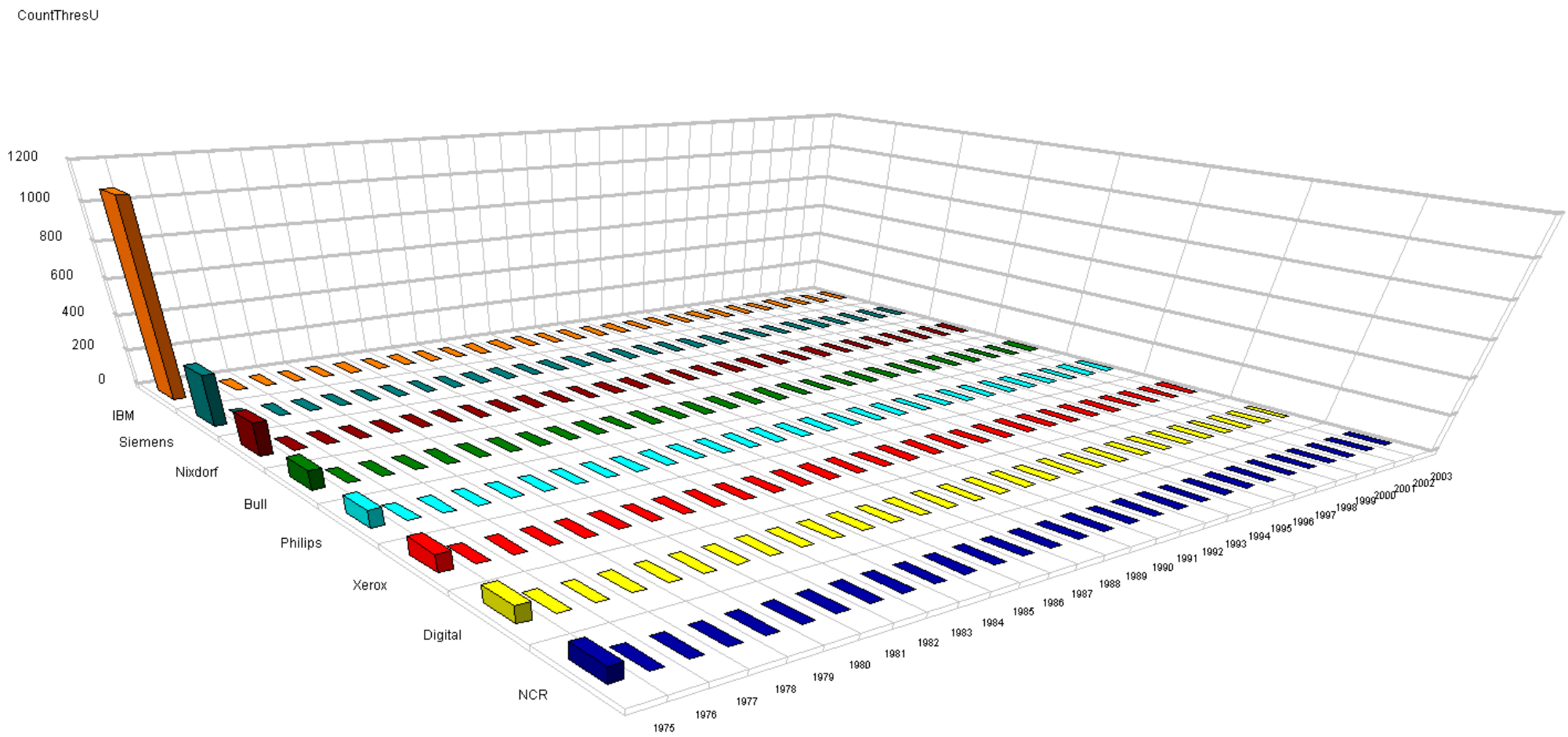


Fig. 120: Visualization of the drill-down dimension “Vendor” for segment “1975” with TRQ threshold filtering

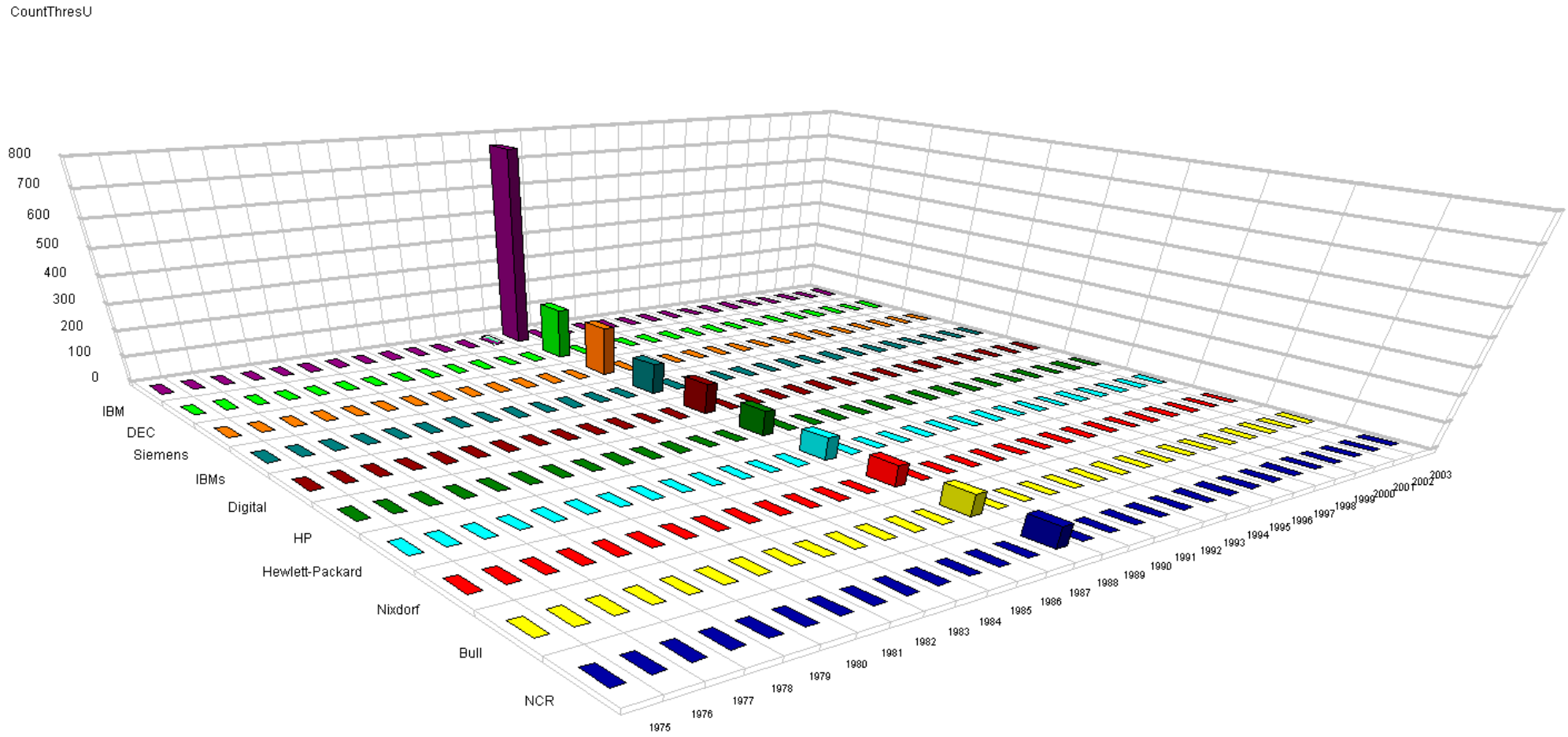


Fig. 121: Visualization of the drill-down dimension “Vendor” for segment “1988” with TRQ threshold filtering

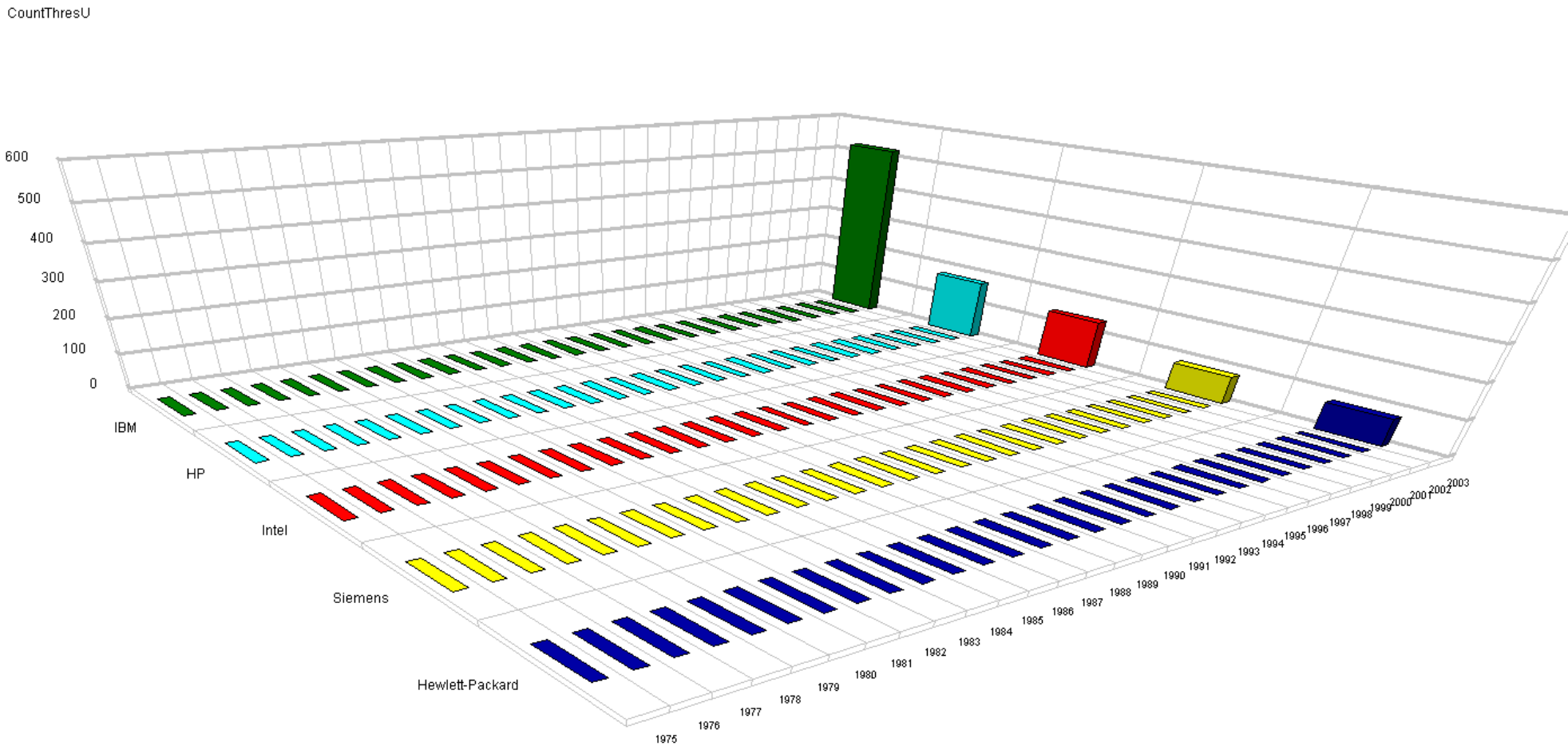


Fig. 122: Visualization of the drill-down dimension “Vendor” for segment “2003” with TRQ threshold filtering

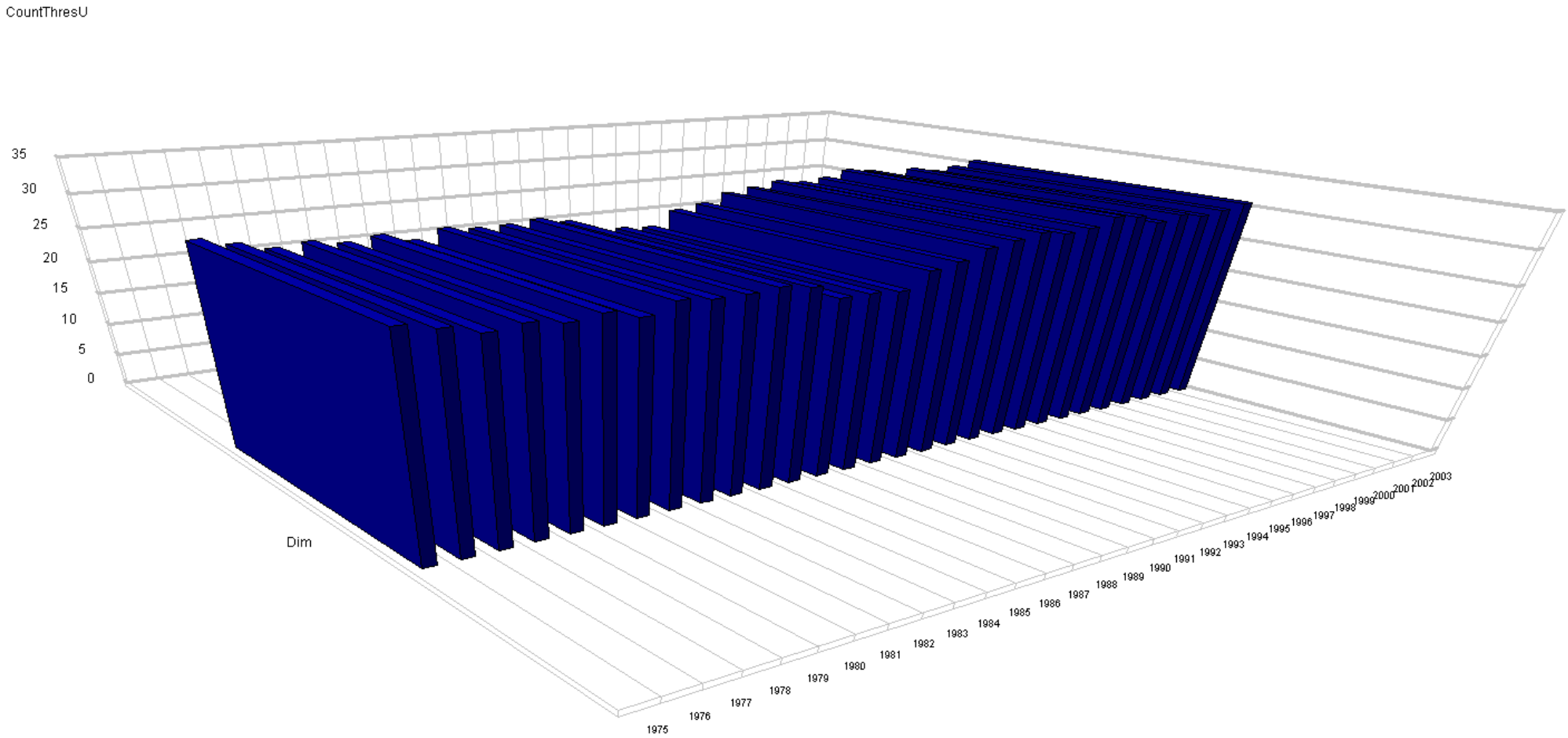


Fig. 123: Visualization of the start node of “Dim” taxonomy for each yearly segment with TRQ threshold filtering for corpus segment C_V of test set CW_{5k}

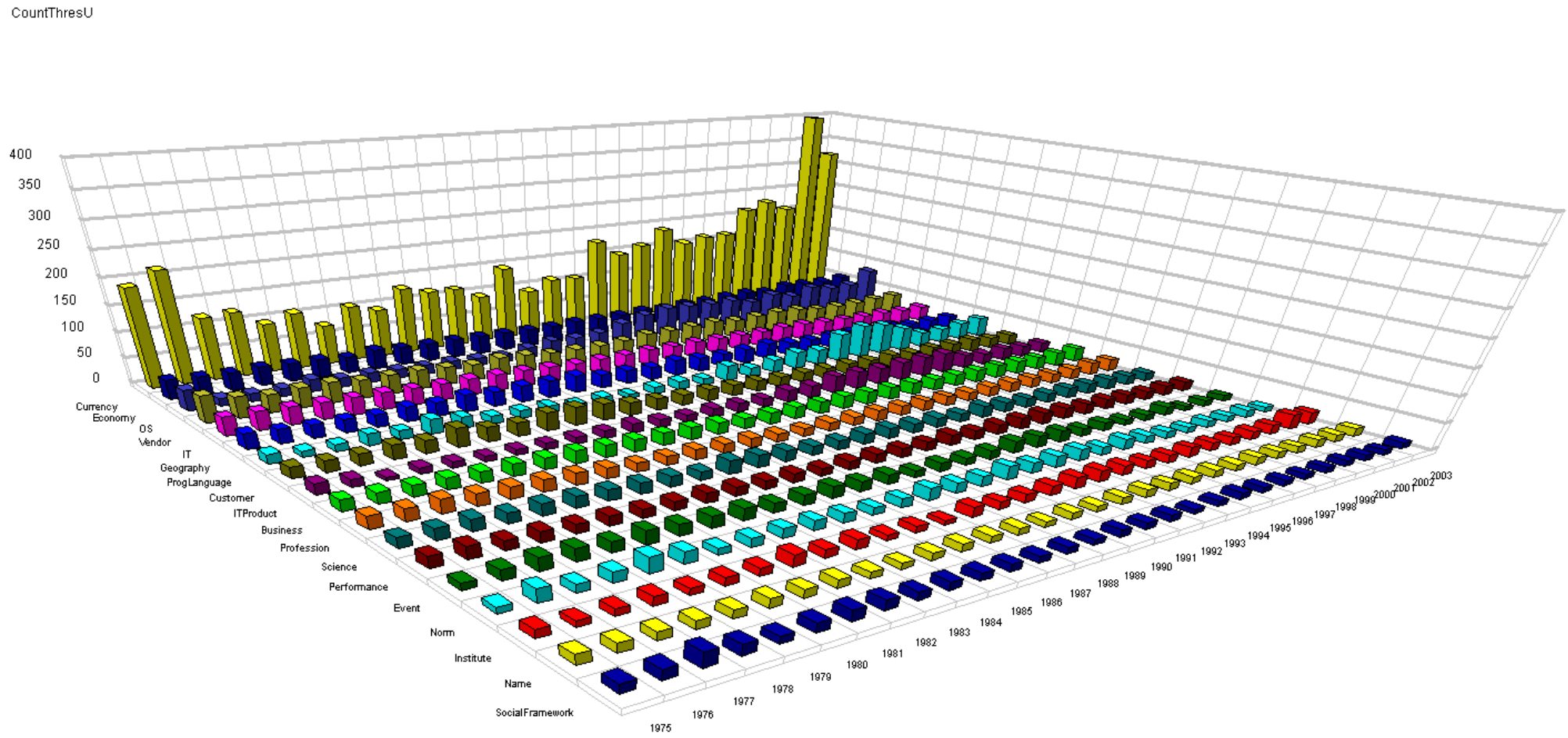


Fig. 124: Visualization of the 2nd level of "Dim" taxonomy for each yearly segment with TRQ threshold filtering for corpus segment C_V of test set CW_{5k}

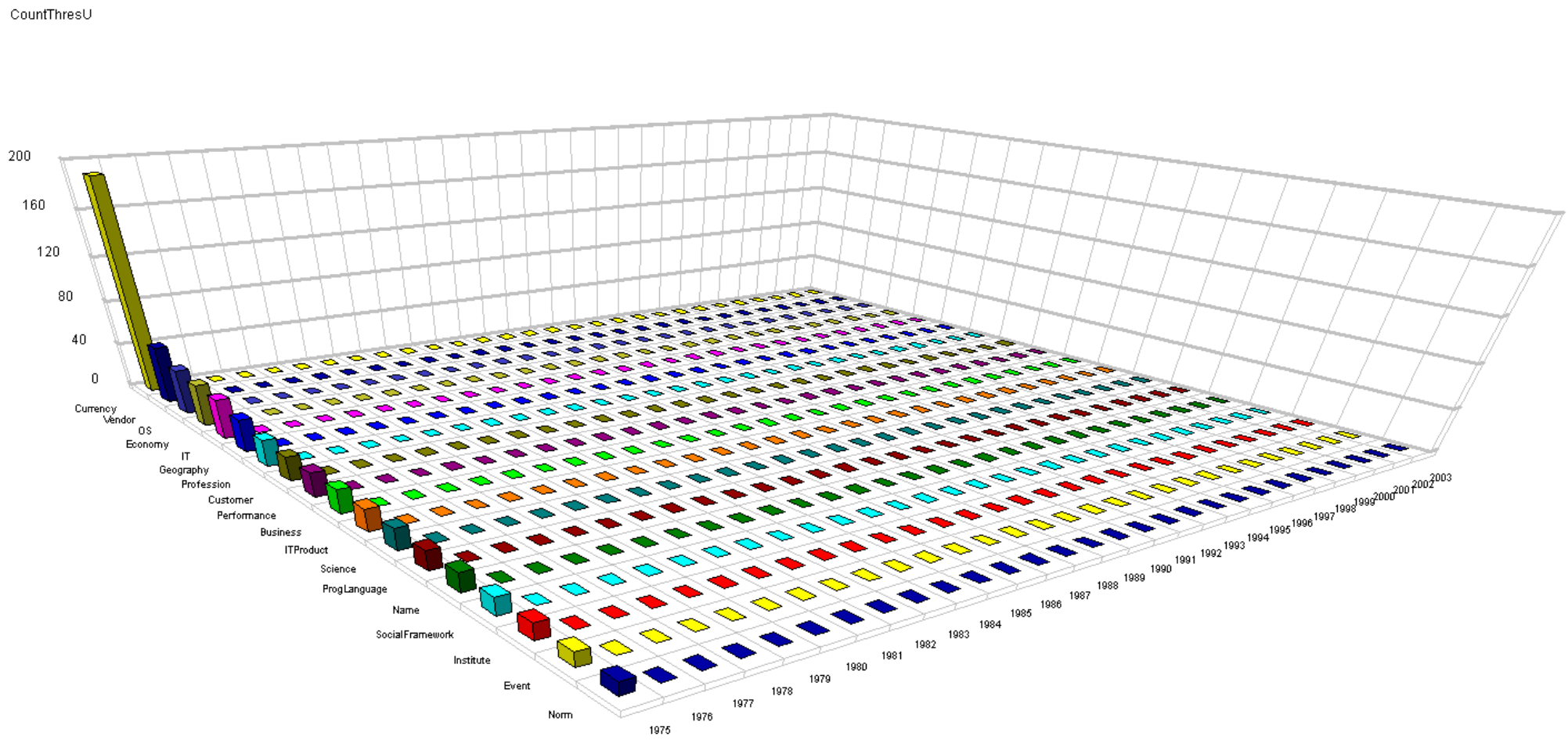


Fig. 125: Visualization of the 2nd level of “Dim” taxonomy for segment “1975” with TRQ threshold filtering for corpus segment C_V of test set CW_{5k}

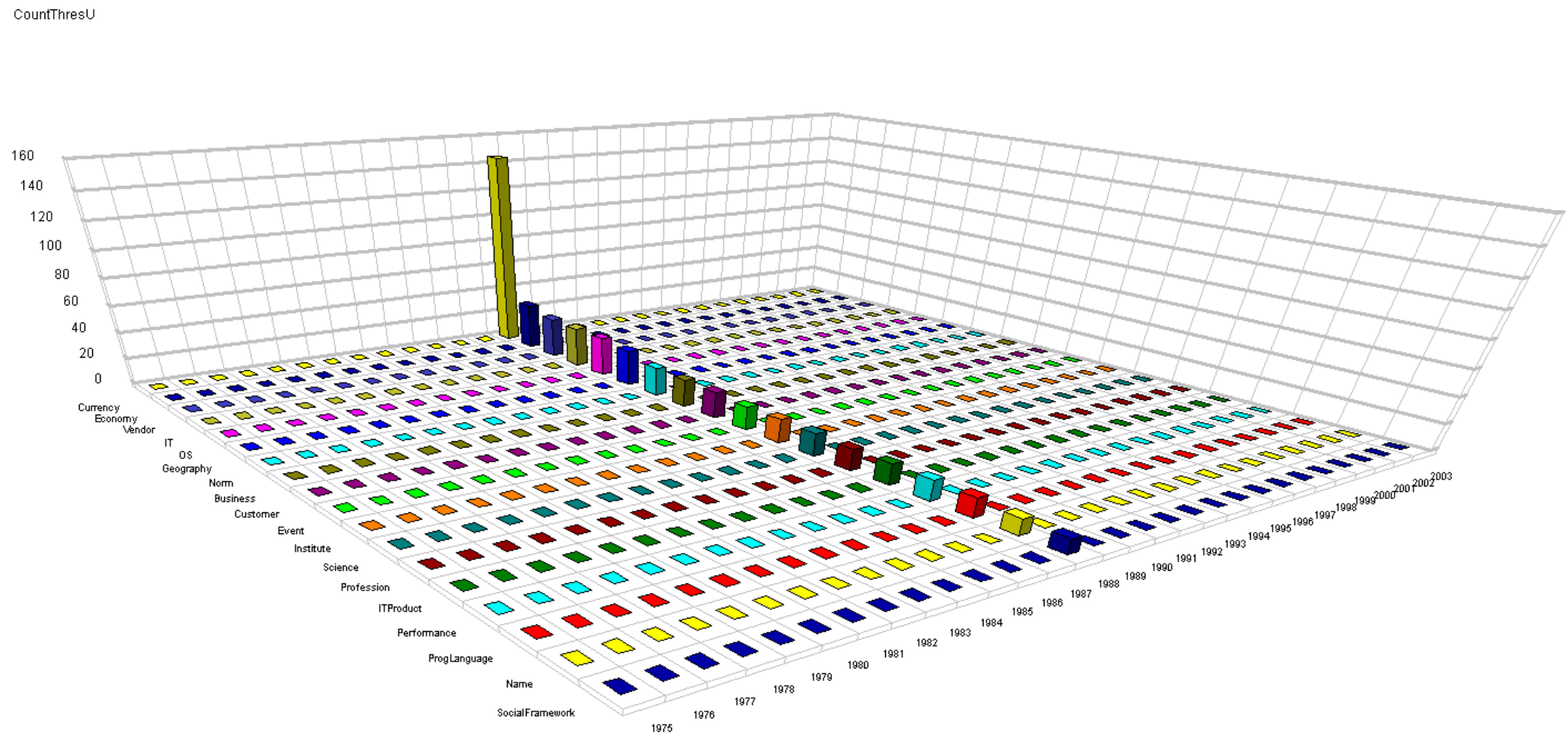


Fig. 126: Visualization of the 2nd level of “Dim” taxonomy for segment “1988” with TRQ threshold filtering for corpus segment C_V of test set CW_{5k}

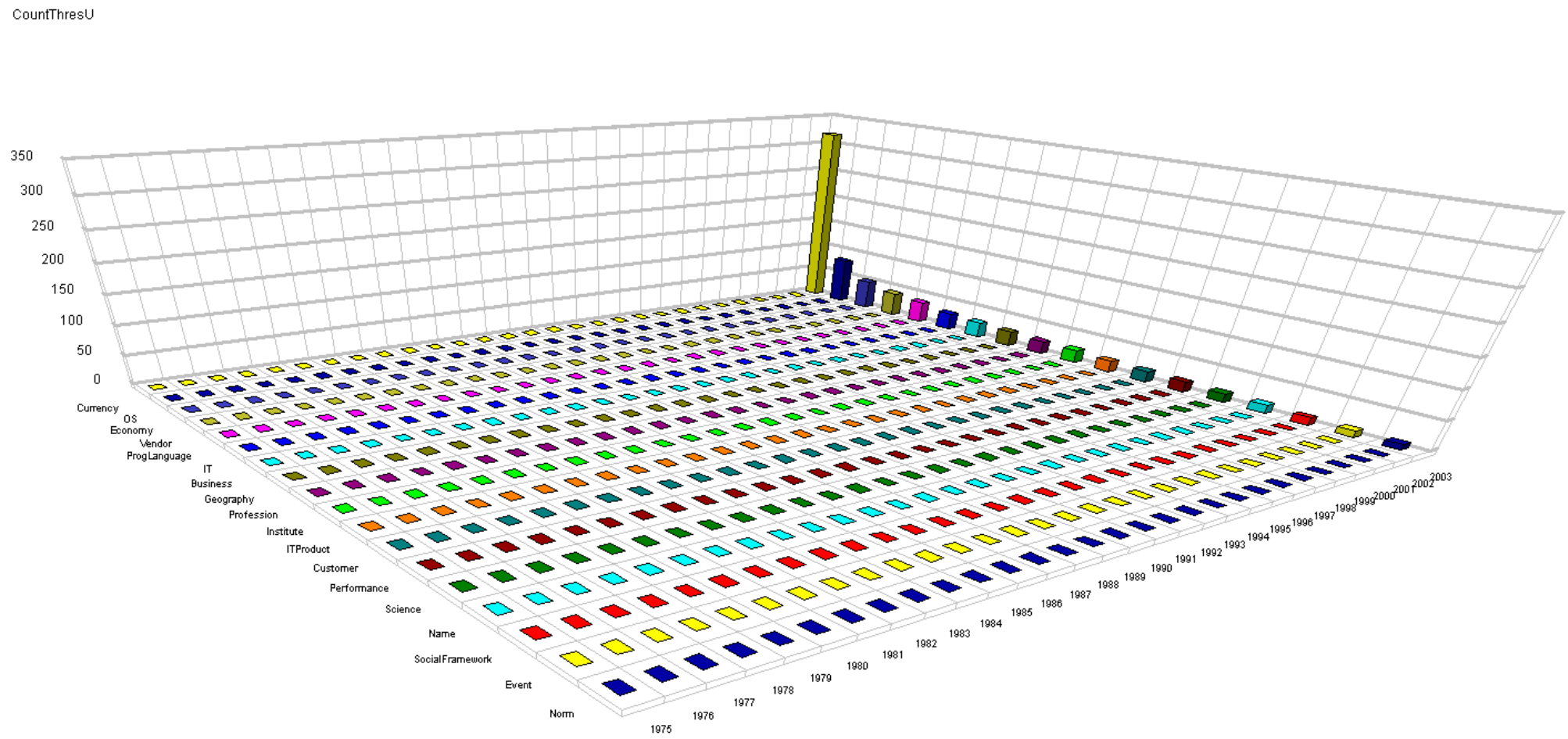


Fig. 127: Visualization of the 2nd level of “Dim” taxonomy for segment “2003” with TRQ threshold filtering for corpus segment C_V of test set CW_{5k}

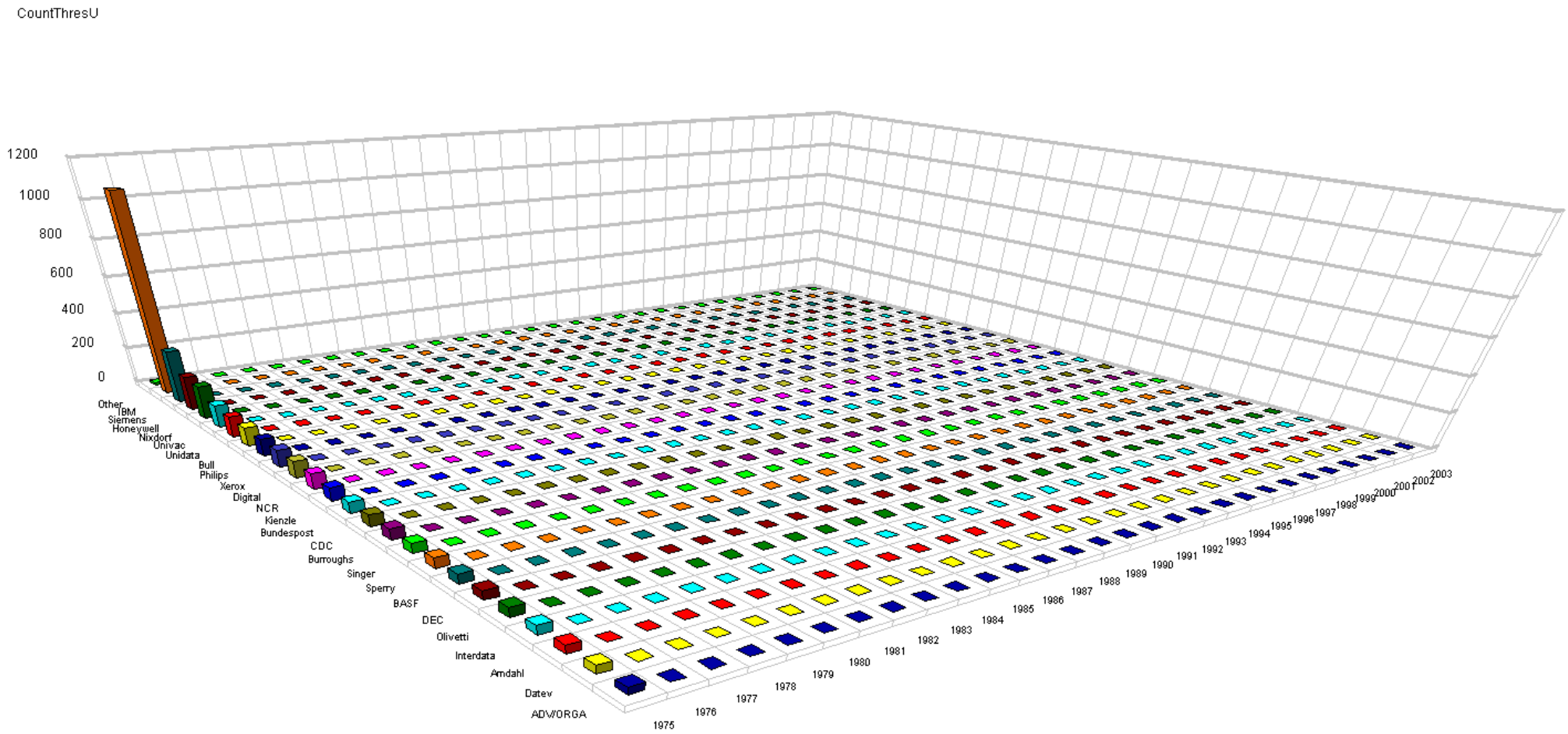


Fig. 128: Visualization of the drill-down dimension “Vendor” for segment “1975” with TRQ threshold filtering

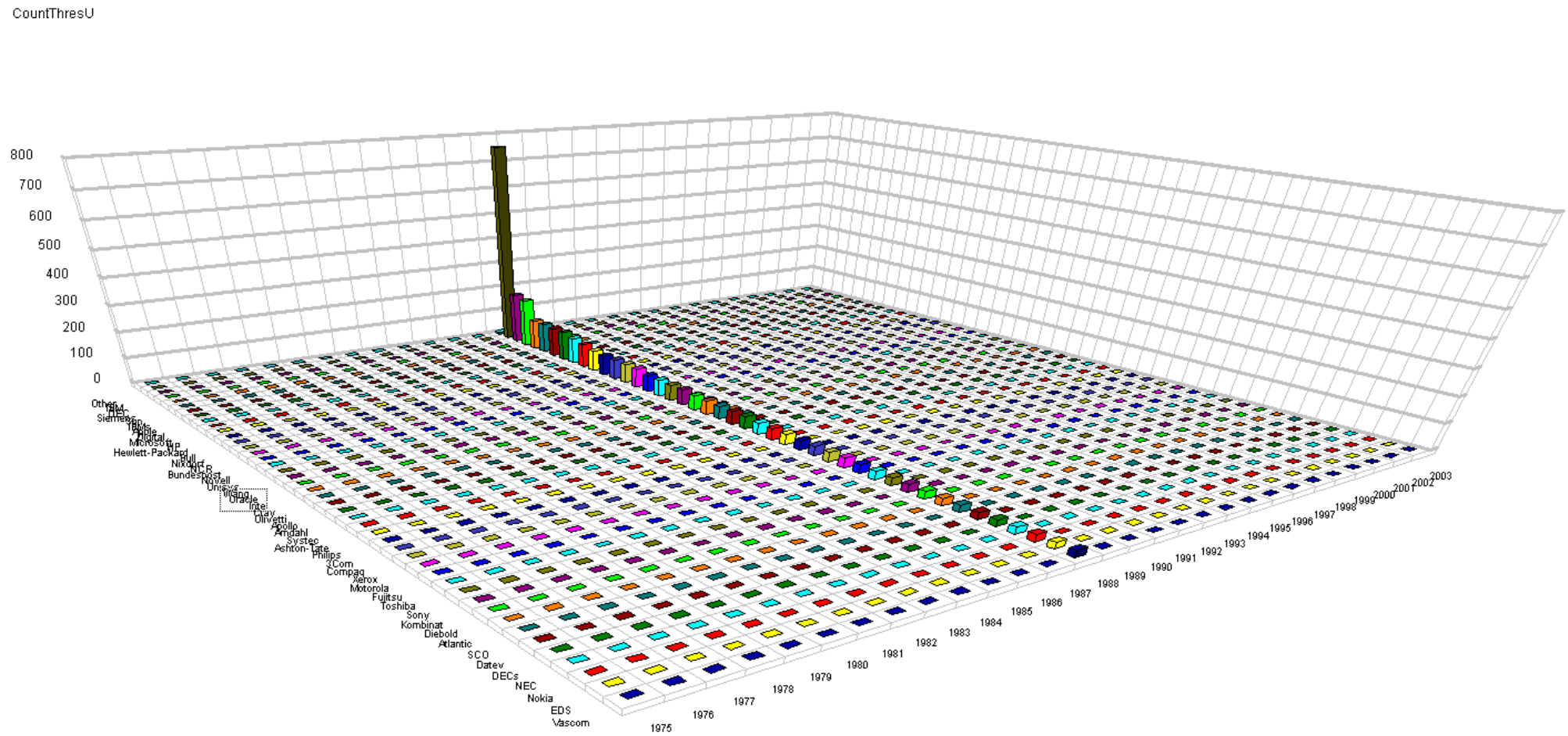


Fig. 129: Visualization of the drill-down dimension “Vendor” for segment “1988” with TRQ threshold filtering

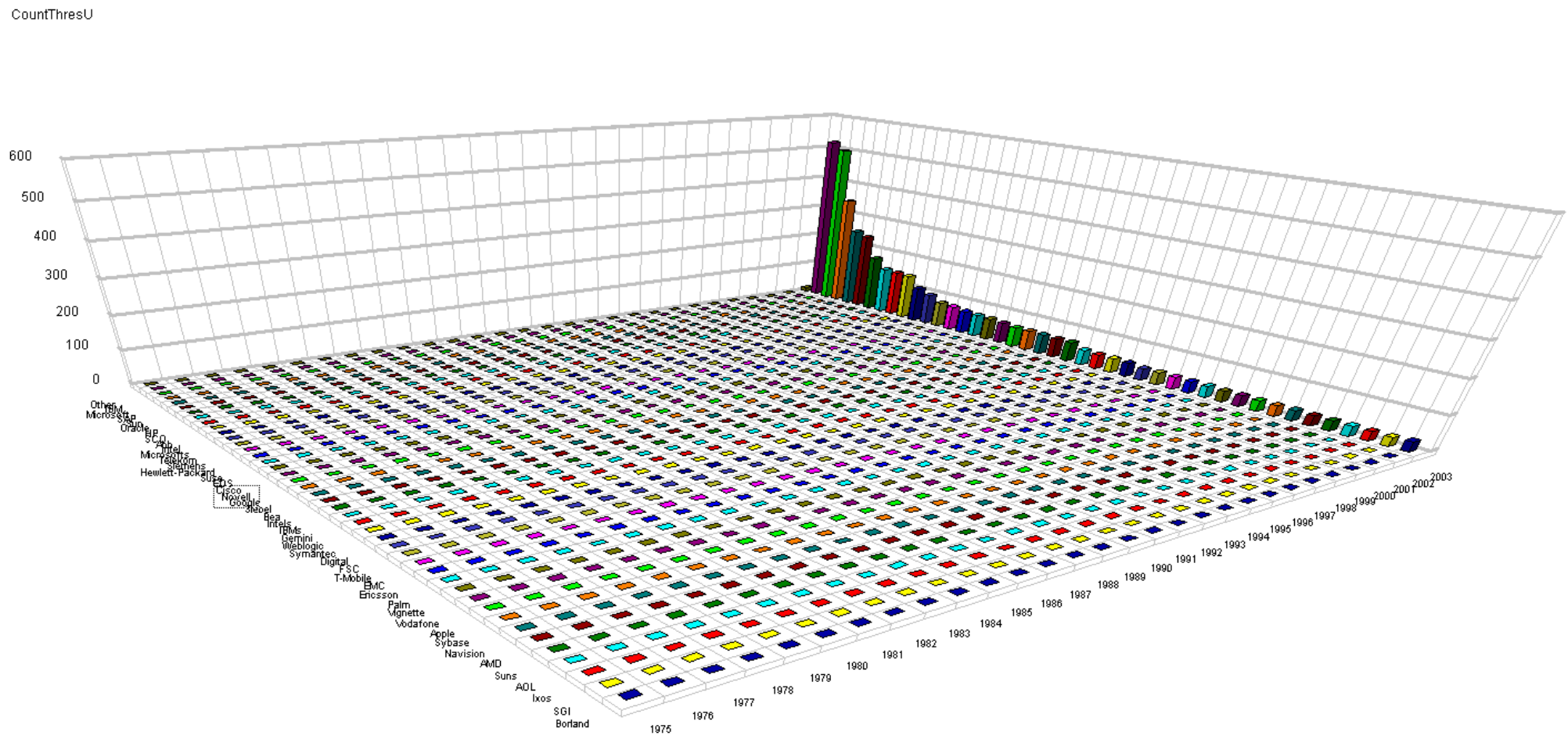


Fig. 130: Visualization of the drill-down dimension “Vendor” for segment “2003” with TRQ threshold filtering

Taxonomy CC_Dim (Allianz, Computerwoche):

Table 68: Taxonomy CC_Dim (Allianz, Computerwoche)

CC_Dim	Term	CC_Dim	Term	CC_Dim	Term
Article	das	Conjunction	und	Preposition	von
Article	Das	Conjunction	Und	Preposition	Von
Article	der	Particle	auch	Preposition	zu
Article	Der	Particle	Auch	Preposition	Zu
Article	des	Preposition	auf	Pronoun	dem
Article	Des	Preposition	Auf	Pronoun	Dem
Article	die	Preposition	bei	Pronoun	den
Article	Die	Preposition	Bei	Pronoun	Den
Article	ein	Preposition	fuer	Pronoun	sich
Article	Ein	Preposition	Fuer	Pronoun	Sich
Article	eine	Preposition	in	Verb	ist
Article	Eine	Preposition	In	Verb	Ist
Article	im	Preposition	mit	Verb	werden
Article	Im	Preposition	Mit	Verb	Werden
Conjunction	als	Preposition	um		
Conjunction	Als	Preposition	Um		

Taxonomy Dim (Computerwoche):

Table 69: Taxonomy Dim (Computerwoche)

Dim	Term	Dim	Term	Dim	Term
Business	Ablauforganisation	Institute	Plaut	IT	Verbindungsdaten
Business	Absatz	Institute	Ploenzke	IT	Verkabelung
Business	Abschreibung	Institute	Pricewaterhouse-Coopers	IT	Vernetzung
Business	Abschreibungen	Institute	PSI	IT	verschluesselt
Business	absetzen	Institute	PWC	IT	Verschluesselung
Business	Abteilung	Institute	SBS	IT	Versenden
Business	Abteilungen	Institute	SCS	IT	Version
Business	Accounting	Institute	Sullivan	IT	Versionen
Business	AGB	Institute	T-Systems	IT	verteilt
Business	Agentur	Institute	Times	IT	verteilte
Business	Agenturen	Institute	vwd	IT	Video
Business	Akquisition	IT	286er	IT	Video-
Business	Akquisitionen	IT	386er	IT	View
Business	Akten	IT	3D	IT	Viren
Business	amortisiert	IT	486er	IT	Virtual
Business	anfertigen	IT	Abarbeiten	IT	virtuelle
Business	angestellt	IT	Abarbeitung	IT	virtuellen
Business	Angestellte	IT	Abfragesprache	IT	Virus
Business	Angestellten	IT	abgespeichert	IT	Vision
Business	Angestellter	IT	Ablaufsteuerung	IT	Visual
Business	Anlagenbuchhaltung	IT	abspeichern	IT	Voice
Business	anschaffen	IT	Abteilungsrechner	IT	VoIP
Business	Anschaffung	IT	AD/Cycle	IT	Volt
Business	Anschaffungskosten	IT	Adabas	IT	VPN
Business	Anschaffungspreis	IT	Adapter	IT	VPNs
Business	Anteile	IT	Addiert	IT	Waehlleitungen
Business	anzuschaffen	IT	Administration	IT	WAP
Business	Arbeitnehmer	IT	Administrative	IT	Warehouse
Business	Arbeitnehmern	IT	administrativen	IT	Wartung
Business	Arbeitnehmers	IT	Adressraum	IT	Web
Business	Arbeitsablaeufe	IT	ADSL	IT	Web-
Business	Arbeitsaufwand	IT	Agent	IT	Web-basierte
Business	Arbeitsbedingungen	IT	Agenten	IT	Web-Browser
Business	Arbeitsbelastung	IT	Algorithmen	IT	Web-Seite
Business	Arbeitsergebnisse	IT	Algorithmus	IT	Web-Seiten
Business	Arbeitserleichterung	IT	Analog	IT	Web-Server
Business	Arbeitsgang	IT	analoge	IT	Web-Services
Business	Arbeitsgebiet	IT	analogen	IT	Web-Site
Business	Arbeitsgebiete	IT	analoger	IT	Web-Sites
Business	Arbeitskraft	IT	Analyse	IT	Website
Business	Arbeitsleistung	IT	Analysen	IT	Websites
Business	Arbeitsmittel	IT	Analysieren	IT	weiterentwickelt
Business	Arbeitsorganisation	IT	analysiert	IT	Weiterentwicklung
Business	Arbeitsplaetze	IT	Analysis	IT	Werkzeug
Business	Arbeitsplaetzen	IT	Analytiker	IT	Wide
Business	Arbeitsplatz	IT	Anbindung	IT	Wireless
Business	Arbeitsplatzes	IT	Anforderung	IT	wissen
Business	Arbeitszeit	IT	Anforderungen	IT	Wissens-Management
Business	Arbeitszeiten	IT	Anforderungskatalog	IT	WLAN
Business	Archiv	IT	anpassen	IT	WLANs
Business	archivieren	IT	Anpassungen	IT	Workflow
Business	archiviert	IT	Anpassungsaufwand	IT	Workflow-
Business	Archivierung	IT	Anruf	IT	Workstation
Business	Aufsichtsrat	IT	Anrufe	IT	Workstations
Business	Aufsichtsrates	IT	anrufen	IT	WWW
Business	Auftraege	IT	ANSI	IT	X-Terminals
Business	Auftrag	IT	Ansteuerung	IT	XML
Business	Auftrags	IT	Anweisung	IT	Zeile
Business	Auftragsabwicklung	IT	Anweisungen	IT	Zeilen
Business	Auftragsbearbeitung	IT	Anwender	IT	Zeilendrucker
Business	Auftragsdaten	IT	Anwenderbericht	IT	Zentraleinheit
Business	Auftragseingang	IT	Anwendergruppen	IT	Zentraleinheiten

Dim	Term	Dim	Term	Dim	Term
Business	Auftragserfassung	IT	Anwendern	IT	Zentralrechner
Business	Auftragsvergabe	IT	Anwenderprogramme	IT	Zoll
Business	Auftragsverwaltung	IT	Anwenders	IT	Zubehoer
Business	Auftragsvolumen	IT	Anwenderseite	IT	Zugangskontrolle
Business	Aufwandskalkulation	IT	Anwendersicht	IT	Zugriff
Business	aufwenden	IT	Anwendersoftware	IT	Zwischenspeicherung
Business	Aufwendungen	IT	Anwendung	ITProduct	80386
Business	Ausgabe	IT	Anwendungen	ITProduct	Amadeus
Business	Ausgaben	IT	Anwendungsentwicklung	ITProduct	Apache
Business	Auslieferung	IT	Anwendungspakete	ITProduct	AS/400
Business	Auslieferungen	IT	Anwendungsprogramm	ITProduct	CII
Business	Ausschreibung	IT	Anwendungsprogramme	ITProduct	CII-HB
Business	Ausschreibungen	IT	Anwendungssoftware	ITProduct	Cii-Honeywell
Business	Aussenstellen	IT	Anwendungssystem	ITProduct	Citrix
Business	Bearbeiter	IT	Anwendungssysteme	ITProduct	CII
Business	beauftragen	IT	Anwendungssystemen	ITProduct	Comet
Business	beauftragt	IT	API	ITProduct	Communicator
Business	beauftragte	IT	APIs	ITProduct	Componentware
Business	Belegschaft	IT	Apparate	ITProduct	CYBER
Business	beliefern	IT	Applets	ITProduct	DB2
Business	beliefert	IT	Appliance	ITProduct	dBase
Business	Berichtswesen	IT	Application	ITProduct	Delphi
Business	Berichtszeitraum	IT	Applications	ITProduct	Domino
Business	beschaffen	IT	Applied	ITProduct	DPS
Business	beschafft	IT	Applikation	ITProduct	DSS
Business	Beschaffung	IT	Applikationen	ITProduct	E-Serie
Business	Bestaende	IT	Applikations-Server	ITProduct	Eclipse
Business	bestellen	IT	APPN	ITProduct	EDI-sys
Business	bestellt	IT	Arbeitsdatei	ITProduct	Epoc
Business	bestellte	IT	Arbeitsplatzcomputer	ITProduct	Excel
Business	bestellten	IT	Arbeitsplatzrechner	ITProduct	Explorer
Business	Bestellung	IT	Arbeitsspeicher	ITProduct	Flash
Business	Bestellungen	IT	Arbeitsstationen	ITProduct	Foxpro
Business	Bestellwesen	IT	Arbeitsvorbereitung	ITProduct	GRASP
Business	Beteiligung	IT	Architekturen	ITProduct	Groupwise
Business	Beteiligungen	IT	Array	ITProduct	H-Serie
Business	Betrieb	IT	ASCII	ITProduct	IBM-kompatiblen
Business	Betriebe	IT	ASP	ITProduct	IBM-Mikros
Business	Betriebes	IT	Assembler	ITProduct	IBM-PC
Business	betriebliche	IT	ATM	ITProduct	IBM-System
Business	betrieblichen	IT	ATM-	ITProduct	IDM
Business	betrieblicher	IT	ATM-Forum	ITProduct	Intranetware
Business	Betriebskosten	IT	Attribute	ITProduct	Itanium
Business	Betriebsrat	IT	aufgezeichnet	ITProduct	Jboss
Business	Betriebswirtschaft	IT	aufzeichnen	ITProduct	Jobstairs
Business	betriebswirtschaftlich	IT	Aufzeichnung	ITProduct	KIS
Business	betriebswirtschaftliche	IT	Aufzeichnungen	ITProduct	Knowledgeware
Business	betriebswirtschaftlichen	IT	Ausbaustufe	ITProduct	LAN-Manager
Business	betriebswirtschaftlicher	IT	Ausfallsicherheit	ITProduct	Lazarus
Business	betriebswirtschaftliches	IT	ausgestattet	ITProduct	Lisa
Business	Bewerber	IT	Auswahlkriterien	ITProduct	Lotus
Business	Bewerbern	IT	Auswertung	ITProduct	LU6
Business	Bewerbungen	IT	Authentifizierung	ITProduct	MAC
Business	Bilanz	IT	Authentisierung	ITProduct	Macintosh
Business	Bilanzierung	IT	Automaten	ITProduct	Manugistics
Business	Board	IT	Automatic	ITProduct	Memmaker
Business	Boards	IT	Automation	ITProduct	Merced
Business	Buchfuehrung	IT	automatisieren	ITProduct	MicroVAX
Business	Buchhaltung	IT	automatisiert	ITProduct	Mikrokanal
Business	Budget	IT	Automatisierte	ITProduct	MPP-Systeme
Business	Budgets	IT	automatisierten	ITProduct	MQ
Business	Buero	IT	automatisierter	ITProduct	Mupid
Business	Buero-	IT	Automatisierung	ITProduct	Mysap
Business	Bueroarbeit	IT	Backbone	ITProduct	MySQL
Business	Bueros	IT	Backup	ITProduct	Navigator
Business	Chairman	IT	Baender	ITProduct	Netview
Business	Chef	IT	Band	ITProduct	Netware
Business	Chefs	IT	Bandbreite	ITProduct	Newton
Business	Controlling	IT	Bandlaufwerke	ITProduct	Notes
Business	Corporation	IT	Basic	ITProduct	Nova
Business	Cost	IT	Basis	ITProduct	NT
Business	Darlehen	IT	Basiskonfiguration	ITProduct	One
Business	Decree	IT	Batch	ITProduct	Openview
Business	Director	IT	Batch-Verarbeitung	ITProduct	Outlook
Business	Direktoren	IT	Bauelemente	ITProduct	Overture
Business	Disposition	IT	Bauelementen	ITProduct	PAC
Business	Distribution	IT	Baugruppe	ITProduct	Pagemaker
Business	Distributor	IT	Baugruppen	ITProduct	PDP
Business	Dividende	IT	BDE	ITProduct	Pentium
Business	Division	IT	Beans	ITProduct	Pentium-Rechner
Business	DV-Abteilung	IT	bedienbar	ITProduct	PET
Business	DV-Abteilungen	IT	Bedienbarkeit	ITProduct	Power-Mac
Business	Ebit	IT	Bediener	ITProduct	Power-PC
Business	Ebitda	IT	Bedienung	ITProduct	Powerbuilder
Business	EDV-Kosten	IT	Befehl	ITProduct	PS/2
Business	Effektivitaet	IT	Befehle	ITProduct	R/2
Business	Eigenentwicklung	IT	Befehlen	ITProduct	R/3
Business	Eigenentwicklungen	IT	Befehlssatz	ITProduct	RS/6000
Business	Eigenkapital	IT	Belegleser	ITProduct	Smartsuite
Business	eingekauft	IT	Beleglesung	ITProduct	Streettalk
Business	eingespart	IT	benutzerfreundlich	ITProduct	Symmetrix
Business	einkalkuliert	IT	benutzerfreundliche	ITProduct	Symphony
Business	Einkauf	IT	benutzerfreundlichen	ITProduct	Tivoli

Dim	Term	Dim	Term	Dim	Term
Business	einkaufen	IT	benutzerfreundlicher	ITProduct	TP-Monitor
Business	Einkaufs	IT	Benutzerfreundlichkeit	ITProduct	Transputer
Business	Einnahmen	IT	Benutzergruppen	ITProduct	UCC
Business	einnehmen	IT	Benutzerinformation	ITProduct	Unicenter
Business	Einsatz	IT	Benutzern	ITProduct	Unify
Business	einsparen	IT	Benutzeroberflaeche	ITProduct	Unixware
Business	Einsparung	IT	Benutzeroberflaechen	ITProduct	VAX
Business	Einsparungen	IT	Benutzers	ITProduct	VAX-Rechnern
Business	einstellen	IT	Benutzerschnittstelle	ITProduct	Ventura
Business	Entgelt	IT	Benutzerschnittstellen	ITProduct	Victor
Business	entlassen	IT	Benutzerservice	ITProduct	Visio
Business	Entlassungen	IT	berechnen	ITProduct	VM/370
Business	Entleiher	IT	berechnet	ITProduct	VS
Business	Entlohnung	IT	Berechnung	ITProduct	Warp
Business	Entwicklungsabteilung	IT	Berechnungen	ITProduct	Webmethods
Business	Entwicklungsaufwand	IT	Beta	ITProduct	Websphere
Business	Entwicklungskosten	IT	Betaversion	ITProduct	Windows-
Business	Entwicklungsprojekte	IT	Betreiber	ITProduct	Windows-Version
Business	Entwicklungsprozess	IT	betriebsbereit	ITProduct	Word
Business	Entwicklungsprozesses	IT	Betriebsbereitschaft	ITProduct	Wordperfect
Business	Entwicklungsstand	IT	Betriebsdaten	ITProduct	Works
Business	Entwicklungsvorhaben	IT	Betriebsdatenerfassung	ITProduct	XA
Business	Entwicklungszentrum	IT	Betriebssicherheit	ITProduct	XP
Business	erfinden	IT	Betriebssystem	ITProduct	XT
Business	Erfinder	IT	Betriebssysteme	ITProduct	Zuse
Business	Erfindung	IT	Betriebssystemen	Name	Adam
Business	Erfindungen	IT	Betriebssystems	Name	Albert
Business	Erfolg	IT	Bewegungsdaten	Name	Alfred
Business	Erfolge	IT	Bewertungskriterien	Name	Andreas
Business	Erfolgs	IT	Bibliothek	Name	Anton
Business	Erforschung	IT	Bibliotheken	Name	Arnd
Business	erfunden	IT	Bilder	Name	Arno
Business	Ersparnis	IT	Bildern	Name	Arnold
Business	Ertrag	IT	Bildes	Name	Arthur
Business	erwirtschaftet	IT	Bildpunkten	Name	Arzubi
Business	Erzeugen	IT	Bildschirm	Name	Augustin
Business	Erzeugnisse	IT	Bildschirm-Terminals	Name	Axel
Business	Erzeugung	IT	Bildschirmarbeitsplaetze	Name	Baeurer
Business	expandierenden	IT	Bildschirmarbeitsplatz	Name	Ballmer
Business	expandiert	IT	Bildschirme	Name	Becker
Business	Expansion	IT	Bildschirmen	Name	Berger
Business	Experten	IT	Bildschirmterminals	Name	Bernd
Business	Fabrik	IT	Bildschirmtext	Name	Bernhard
Business	Fabriken	IT	Bit	Name	Bill
Business	Fachabteilung	IT	Bits	Name	Bluemmel
Business	Fachabteilungen	IT	Blades	Name	Blumenthal
Business	Fachbereich	IT	Bluetooth	Name	Bob
Business	Fachbereiche	IT	BMC	Name	Boesenberg
Business	Fachbereichen	IT	box	Name	Bojanowsky
Business	Fachbereichs	IT	Bridge	Name	Brillinger
Business	Facilities	IT	Brief	Name	Butler
Business	Fakturierung	IT	Briefe	Name	Carl
Business	fertigen	IT	Briefen	Name	Charles
Business	Fertigstellung	IT	Broker	Name	Christian
Business	fertigt	IT	Browser	Name	Christoph
Business	Fertigung	IT	Btx	Name	Clara
Business	Fertigungsbereich	IT	Btx-	Name	Claus
Business	Filiale	IT	Btx-System	Name	David
Business	Filialen	IT	Buchstaben	Name	Dieter
Business	Finanzbuchhaltung	IT	Bueroautomation	Name	Dietrich
Business	Finanzchef	IT	Buerocomputer	Name	Drodofsky
Business	finanzieren	IT	Buerokommunikation	Name	Eberhard
Business	finanziert	IT	Bueromaschinen	Name	Eckhard
Business	Finanzierung	IT	Buerotechnik	Name	Edward
Business	Finanzkraft	IT	Business-TV	Name	Edwards
Business	Finanzplanung	IT	Byte	Name	Ellison
Business	Finanzwesen	IT	Bytes	Name	Erich
Business	Firmenchef	IT	CA-Sort	Name	Erwin
Business	firmeneigenen	IT	Cache	Name	Fiorina
Business	Firmengruppe	IT	CAD	Name	Frank
Business	firmeninterne	IT	CAD/CAM	Name	Frankenberg
Business	Firmennamen	IT	Call-Center	Name	Franz
Business	firmieren	IT	CAM	Name	Friedrich
Business	firmiert	IT	Card	Name	Fritz
Business	Fiskaljahr	IT	Carrier	Name	Ganzhorn
Business	Fiskalquartal	IT	CASE	Name	Gardner
Business	Fortschreibung	IT	CASE-Tools	Name	Geis
Business	Fuehrung	IT	CBT	Name	Geisler
Business	Fuehrungskraefte	IT	CD	Name	Georg
Business	Funktionsbereiche	IT	CD-ROM	Name	George
Business	Funktionsbereichen	IT	CD-ROM-Laufwerk	Name	Gerald
Business	Gebuehren	IT	CD-ROMs	Name	Gerd
Business	gefertigt	IT	CDs	Name	Gerhard
Business	gegruendet	IT	Chain	Name	Gerstner
Business	gegruendete	IT	Channel	Name	Gert
Business	gegruendeten	IT	Chassis	Name	Grove
Business	Gehaelter	IT	check	Name	Guembel
Business	Gehalt	IT	Checkliste	Name	Guenter
Business	Gehaltsabrechnung	IT	Checklisten	Name	Gupta
Business	Geheimhaltung	IT	Chip	Name	Hans
Business	Geheimnis	IT	Chip-Karte	Name	Hans-Dieter
Business	gekostet	IT	Chips	Name	Hans-Joachim
Business	geliefert	IT	CICS	Name	Harald

Dim	Term	Dim	Term	Dim	Term
Business	gelieferte	IT	CIM	Name	Harry
Business	gelieferten	IT	Cincom	Name	Hartmut
Business	gemietet	IT	Circuit	Name	Hauff
Business	geordert	IT	Class	Name	Heinrich
Business	Gesamtaufwand	IT	Client	Name	Heinz
Business	Gesamtkosten	IT	Client-	Name	Helmut
Business	Gesamtumsatz	IT	Client-Server	Name	Henkel
Business	Gesamtumsatzes	IT	Client-Server-	Name	Henning
Business	Geschaeftsbereiche	IT	Client-Server-Architektur	Name	Herbert
Business	Geschaeftsmodell	IT	Client-Server-Architekturen	Name	Hermann
Business	Geschaeftsquartal	IT	Clients	Name	Hoepfner
Business	gespart	IT	Cluster	Name	Hoffmann
Business	Gewaehrleistung	IT	CMS	Name	Horst
Business	Gewerbe	IT	Cobol	Name	Huber
Business	gewerbliche	IT	Cobol-Programme	Name	Hubert
Business	gewerblichen	IT	Code	Name	Huebner
Business	Gewinn	IT	Codes	Name	Joerg
Business	Gewinne	IT	codieren	Name	John
Business	Gewinnwarnung	IT	codiert	Name	Juergen
Business	gezahlt	IT	codierten	Name	Kahn
Business	Gruenden	IT	Codierung	Name	Karl
Business	Gruender	IT	Collaborative	Name	Klaus
Business	Gruendung	IT	Commerce	Name	Kratz
Business	Grundkapital	IT	Community	Name	Kreuter
Business	Haendler	IT	Compiler	Name	Kurt
Business	Handelsunternehmen	IT	Compilern	Name	Larry
Business	handelt	IT	Component	Name	Liebich
Business	handelte	IT	Computer	Name	Lorenz
Business	Handwerk	IT	Computern	Name	Manfred
Business	Hardwareherstellern	IT	Computers	Name	Martin
Business	Hardwarekosten	IT	Computersystem	Name	Matthoefer
Business	Hauptabteilung	IT	Computersysteme	Name	Maurer
Business	Hauptsitz	IT	Computersystemen	Name	McNealy
Business	Hauptversammlung	IT	Computertechnik	Name	Meier
Business	Hauptverwaltung	IT	Computerwelt	Name	Meiko
Business	hergestellt	IT	Computing	Name	Menzel
Business	hergestellten	IT	Concurrent	Name	Meyer
Business	herstellen	IT	Config	Name	Michael
Business	herstellt	IT	Content	Name	Morris
Business	Herstellung	IT	Content-Management	Name	Mueller
Business	Herstellungskosten	IT	Controller	Name	Name
Business	herzustellen	IT	core	Name	Neumueller
Business	Holding	IT	CPU	Name	Oesterle
Business	Informations-Management	IT	CPUs	Name	Ohmen
Business	Inhaber	IT	CRM	Name	Olsen
Business	innerbetriebliche	IT	CRM-	Name	Otto
Business	innerbetrieblichen	IT	Customer-Relationship-Management	Name	Paul
Business	Insider	IT	Data	Name	Peps
Business	investieren	IT	Data-Warehouse	Name	Peter
Business	investiert	IT	Data-Warehousing	Name	Pfeiffer
Business	Investition	IT	Database	Name	Rauser
Business	Investitionen	IT	Datei	Name	Reiss
Business	Investment	IT	Dateien	Name	Ricardo
Business	IT-Abteilung	IT	Daten	Name	Richard
Business	IT-Abteilungen	IT	Datenanalyse	Name	Ricke
Business	IT-Ausgaben	IT	Datenaufbereitung	Name	Robert
Business	IT-Budgets	IT	Datenaustausch	Name	Rolf
Business	IT-Investitionen	IT	Datenbank	Name	Rollins
Business	IT-Kosten	IT	Datenbank-	Name	Runtagh
Business	Kalkulation	IT	Datenbanken	Name	Samenuk
Business	Kapazitaetsplanung	IT	Datenbanksoftware	Name	Samwer
Business	Kapital	IT	Datenbanksystem	Name	Sanders
Business	Kassen	IT	Datenbanksysteme	Name	Scheer
Business	Kaufsache	IT	Datenbanksystemen	Name	Schmidt
Business	Kennzahlen	IT	Datenbanksystems	Name	Schneider
Business	Kerngeschaeft	IT	Datenbasis	Name	SCHOLZ
Business	Kollegen	IT	Datenbestaende	Name	Schueler
Business	Konditionen	IT	Datenbestand	Name	Schulmeyer
Business	Konsolidierung	IT	Dateneingabe	Name	Schwarz-Schilling
Business	Konstruktion	IT	Datenerfassung	Name	Sebastian
Business	Konten	IT	Datenfelder	Name	Sellmer
Business	Konzern	IT	Datenfernuebertragung	Name	Skcurd
Business	Konzerne	IT	Datenfernverarbeitung	Name	Spitschka
Business	Konzerns	IT	Datenhaltung	Name	Stefan
Business	Kooperationen	IT	Datenkommunikation	Name	Stephan
Business	Kostendruck	IT	Datenmenge	Name	Steve
Business	Kostenrechnung	IT	Datenmengen	Name	Stolorz
Business	Kuendigung	IT	Datenmodell	Name	Strack-Zimmermann
Business	Kundenbindung	IT	Datennetz	Name	Stultitia
Business	Lager	IT	Datennetze	Name	Thepot
Business	Lean	IT	Datennetzen	Name	Thomas
Business	Lieferung	IT	Datensammelsystem	Name	Tolkmit
Business	Lizenz	IT	Datensammelsysteme	Name	Trauerwein
Business	Lohn-	IT	Datensammelsystemen	Name	Tsaoussis
Business	Management	IT	Datensatz	Name	Ulrich
Business	Management-	IT	Datensicherheit	Name	Unger
Business	Managements	IT	Datensicherung	Name	Uwe
Business	Manager	IT	Datensichtgeraet	Name	Voss
Business	Manufacturing	IT	Datensichtgeraete	Name	Vucins
Business	Margen	IT	Datenspeicher	Name	Walter
Business	Marketing	IT	Datenspeicherung	Name	Weber
Business	Marketing-	IT	Datenstation	Name	Wedell

Dim	Term	Dim	Term	Dim	Term
Business	Maschinen	IT	Datenstationen	Name	Wellenreuther
Business	Material	IT	Datenstruktur	Name	Wellhoener
Business	Materialwirtschaft	IT	Datenstrukturen	Name	Werner
Business	Miete	IT	Datensysteme	Name	White
Business	Mietsache	IT	Datentechnik	Name	Wiechers
Business	Mitarbeiter	IT	Datentraeger	Name	William
Business	Mitarbeitern	IT	Datentransfer	Name	Winfried
Business	Mitbewerber	IT	Datentypen	Name	Witte
Business	Mobilitaet	IT	Datenuebermittlung	Name	Wolfgang
Business	Motivation	IT	Datenuebertragung	Name	Wolpers
Business	Muttergesellschaft	IT	Datenverarbeitung	Norm	CAPI
Business	Nettogewinn	IT	Datenverarbeitungsanlagen	Norm	CAT
Business	Nettoverlust	IT	Datenverkehr	Norm	CCITT
Business	Neuentwicklung	IT	Datenverwaltung	Norm	CGI
Business	Niederlassung	IT	Datenzugriff	Norm	CMC
Business	Niederlassungen	IT	Datex-P	Norm	CMIP
Business	Obligo	IT	Datumsumstellung	Norm	Corba
Business	Office	IT	Dauerbetrieb	Norm	COSE
Business	organisatorisch	IT	DBMS	Norm	DCE
Business	organisatorische	IT	DDs	Norm	DCOM
Business	organisatorischen	IT	dedizierte	Norm	DD/DS
Business	organisiert	IT	dedizierten	Norm	DIN
Business	Partners	IT	Desktop	Norm	DME
Business	Personal	IT	Desktop-Publishing	Norm	dpi
Business	Personalabrechnung	IT	Desktops	Norm	DSL
Business	Personalentwicklung	IT	Development	Norm	EAI
Business	Personalkosten	IT	Devices	Norm	EAN
Business	Personalwesen	IT	DFUE	Norm	EDI
Business	Pilotprojekt	IT	Dialogverarbeitung	Norm	ENX
Business	Pleite	IT	Dictionary	Norm	Ethernet-
Business	Postfach	IT	Dietz	Norm	FDDI
Business	Preis-/Leistungsverhaeltnis	IT	digitale	Norm	FTAM
Business	Preis-Leistungs-Verhaeltnis	IT	digitalen	Norm	ftp
Business	Pressekonferenz	IT	digitaler	Norm	genormte
Business	Problemloesung	IT	digitales	Norm	GPRS
Business	Problemloesungen	IT	digitalisierte	Norm	HTML
Business	Product	IT	Digitalisierung	Norm	HTML-
Business	Products	IT	Directory	Norm	Hypertext
Business	Produkt	IT	disk	Norm	IBC
Business	Produkt-Telegramme	IT	Diskette	Norm	IEEE
Business	Produkte	IT	Disketten	Norm	Industriestandard
Business	Produktentwicklung	IT	Display	Norm	iSCSI
Business	Produktes	IT	displays	Norm	ISDN-
Business	Produktion	IT	Distributoren	Norm	ISO
Business	Produktivitaet	IT	DMS	Norm	Jini
Business	Produktpalette	IT	Dokument	Norm	LDAP
Business	Produkts	IT	Dokumentation	Norm	MAP
Business	produziert	IT	Dokumentationen	Norm	MMS
Business	Profit	IT	Dokumentationssystem	Norm	Motif
Business	profitabel	IT	Dokumente	Norm	NDS
Business	Project	IT	Dokumenten	Norm	Norm
Business	Projekt	IT	Dokumenten-Management	Norm	Norm-Entwurf
Business	Projekt-Management	IT	Download	Norm	Normen
Business	Projekte	IT	Downsizing	Norm	Normierte
Business	Projekten	IT	DP	Norm	Normung
Business	Projektes	IT	drahtlose	Norm	ODBC
Business	Projektgruppe	IT	drahtlosen	Norm	OMG
Business	Projektkosten	IT	Drive	Norm	Opendoc
Business	Projektmanagement	IT	Druckausgabe	Norm	OSF/1
Business	Projektorganisation	IT	Drucken	Norm	OSF/Motif
Business	Projektplanung	IT	Drucker	Norm	OSI
Business	Projekts	IT	Druckern	Norm	PCI
Business	Projektteam	IT	druckt	Norm	PCI-Bus
Business	Prozesse	IT	DTP	Norm	Postscript
Business	Prozessen	IT	DV	Norm	proprietaere
Business	Qualitaetsfoerderung	IT	DV-	Norm	proprietaeren
Business	Qualitaetszirkel	IT	DV-Anlagen	Norm	Rambus
Business	Rationalisierung	IT	DV-Einsatz	Norm	RMI
Business	Rechnung	IT	DV-System	Norm	SNA
Business	Rechnungswesen	IT	DV-Systeme	Norm	SNMP
Business	Rechnungswesens	IT	DV-Systemen	Norm	Soap
Business	Reingewinn	IT	E-Business	Norm	SWIFT
Business	Reorganisation	IT	e-commerce	Norm	TCP/IP
Business	Revision	IT	E-Government	Norm	Token-Ring
Business	Risiken	IT	E-Learning	Norm	UDDI
Business	Risiko	IT	E-Mail	Norm	V.3
Business	RoI	IT	E-Mails	Norm	V.4
Business	Rolle	IT	E-Plus	Norm	VSE/ESA
Business	RZ	IT	E-Procurement	Norm	VTAM
Business	Schreibtisch	IT	Echtzeit	Norm	X.25
Business	Shop	IT	Edifact	Norm	X.400
Business	Sitz	IT	Editor	Norm	X/Open
Business	SLAs	IT	EDV	OS	AIX
Business	Sparte	IT	EDV-	OS	BS
Business	Sprecher	IT	EDV-Abteilung	OS	BS1000
Business	Standorten	IT	EDV-Anlage	OS	BS2000
Business	Strategie	IT	EDV-Anlagen	OS	CP/M
Business	Strategien	IT	EDV-Bereich	OS	CP/M-86
Business	strategischen	IT	EDV-Hersteller	OS	Daytona
Business	Stueck	IT	EDV-System	OS	Disoss
Business	Stueckzahlen	IT	EDV-Systeme	OS	DOS
Business	Taetigkeit	IT	Einbindung	OS	DOS-
Business	Taetigkeiten	IT	Einfuehrung	OS	DOS/VS

Dim	Term	Dim	Term	Dim	Term
Business	Team	IT	Eingabe	OS	DOS/VSE
Business	Teams	IT	Eingaben	OS	HP-UX
Business	Tochtergesellschaft	IT	eingebunden	OS	Linux
Business	Top-Management	IT	eingelesen	OS	Mac-OS
Business	UDM	IT	Einlesen	OS	MS-DOS
Business	Umsaetze	IT	Einstiegsmodell	OS	MS-Windows
Business	Umsatz	IT	Eintippen	OS	MVS
Business	Umsatzes	IT	EIS	OS	MVS/ESA
Business	Umstrukturierung	IT	Electric	OS	MVS/XA
Business	Unternehmens	IT	Electronic	OS	Nextstep
Business	Unternehmensangaben	IT	Electronics	OS	OS
Business	Unternehmenskultur	IT	elektrische	OS	OS/2
Business	Venture	IT	elektrischen	OS	Sinix
Business	Verantwortung	IT	Elektronik	OS	Solaris
Business	verdient	IT	elektronisch	OS	Symbian
Business	Verguetung	IT	elektronischen	OS	Tru-64-Unix
Business	verkauften	IT	elektronischer	OS	Ultrix
Business	Verlust	IT	Elektronisches	OS	Unix
Business	Verluste	IT	Element	OS	Unix-
Business	vermarkten	IT	Elemente	OS	Unix-System
Business	vermarktet	IT	Elementen	OS	Unix-Systeme
Business	Vermarktung	IT	Empfaenger	OS	Unix-Systemen
Business	Versand	IT	Emulation	OS	Vines
Business	versichert	IT	Emulatoren	OS	VM
Business	Versteigerer	IT	emuliert	OS	VMS
Business	Vertraege	IT	Endgeraete	OS	VSE
Business	Vertrag	IT	Engine	OS	Windows
Business	Vertrages	IT	Engineering	OS	Xenix
Business	Vertragsbedingungen	IT	Enterprise	Performance	Antwortzeit
Business	Vertragspartner	IT	Entity	Performance	Antwortzeiten
Business	vertreiben	IT	Entscheidungstabellen	Performance	Ausfall
Business	vertreibt	IT	Entwicklungsarbeit	Performance	ausfallen
Business	Vertreter	IT	Entwicklungsarbeiten	Performance	Ausfallzeiten
Business	Vertrieb	IT	Entwicklungsphase	Performance	ausgefallen
Business	Vertrieben	IT	Entwicklungsumgebung	Performance	Auslastung
Business	Vertriebs	IT	Entwicklungswerkzeuge	Performance	Baud
Business	Vertriebs-	IT	Equipment	Performance	Benchmark
Business	Vertriebs-Tochter	IT	Erfassen	Performance	Benchmark-Tests
Business	Verwaltungen	IT	Erfassung	Performance	beschleunigen
Business	Vorgangsbearbeitung	IT	Ergonomie	Performance	beschleunigt
Business	Wartungskosten	IT	ergonomischen	Performance	beschleunigte
Business	Werbung	IT	ERP	Performance	Beschleunigung
Business	Werk	IT	ERP-	Performance	BPI
Business	Werke	IT	ERP-System	Performance	Computerleistung
Business	Werkzeugen	IT	Errechnen	Performance	CPU-Zeit
Business	Werkzeugmaschinen	IT	errechnet	Performance	Datendurchsatz
Business	Workgroups	IT	Errechnete	Performance	Datenrate
Business	Zahlung	IT	Ersatzteile	Performance	Datenvolumen
Business	Ziel	IT	Erweiterung	Performance	Defekt
Currency	Cent	IT	Erweiterungen	Performance	defekte
Currency	Cents	IT	ESS	Performance	Druckgeschwindigkeit
Currency	DM	IT	Ethernet	Performance	Durchlaufzeiten
Currency	Dollar	IT	Etikett	Performance	Durchsatz
Currency	ECU	IT	Etiketten	Performance	Einarbeitungszeit
Currency	Euro	IT	Evaluation	Performance	einsatzbereit
Currency	Franc	IT	Evolution	Performance	Entwicklungszeit
Currency	Francs	IT	Exchange	Performance	ergonomische
Currency	Mark	IT	Expertensystem	Performance	Erleichterung
Currency	Pfennig	IT	Expertensysteme	Performance	fehlerfrei
Currency	Pfund	IT	Expertensystemen	Performance	Fehlerrate
Currency	Schilling	IT	Fax	Performance	Folgekosten
Currency	Sterling	IT	feature	Performance	GB
Currency	US-Dollar	IT	Features	Performance	Genauigkeit
Currency	Yen	IT	Fehlerbehandlung	Performance	Geschwindigkeit
Customer	Allianz	IT	Fehlererkennung	Performance	Geschwindigkeiten
Customer	Audi	IT	Fehlerfall	Performance	Gigabit
Customer	Benutzer	IT	Fehlermeldungen	Performance	Gigabyte
Customer	Bundesbahn	IT	Fenster	Performance	Gigahertz
Customer	Bundeswehr	IT	Fenstern	Performance	Gutachten
Customer	Burda	IT	Fernsehen	Performance	Handling
Customer	BVB	IT	Fernwartung	Performance	Hauptspeicherkapazitaet
Customer	BVG	IT	Fertigungssteuerung	Performance	hz
Customer	Ciba	IT	Fertigungstechnik	Performance	Informationsgehalt
Customer	Daimler-Benz	IT	Festplatte	Performance	Kapazitaet
Customer	Daimler-Chrysler	IT	Festplatten	Performance	Kapazitaeten
Customer	Dekra	IT	Fibre	Performance	Kbit/s
Customer	DeMoulas	IT	File	Performance	Laufzeit
Customer	DMV	IT	files	Performance	Laufzeiten
Customer	DV-Anwender	IT	Film	Performance	Leistungsfaehigkeit
Customer	EAM	IT	Filter	Performance	Leistungssteigerung
Customer	Ebay	IT	Firewall	Performance	MB/s
Customer	EDV-Anwender	IT	Firewalls	Performance	Mbit/s
Customer	Endanwender	IT	Firmware	Performance	Megabit
Customer	Endbenutzer	IT	Floppy	Performance	Megabyte
Customer	Endkunden	IT	Folien	Performance	Megahertz
Customer	Erstanwender	IT	Format	Performance	Metriken
Customer	Flughafen	IT	Formate	Performance	Mhz
Customer	Ford	IT	Formaten	Performance	Millisekunden
Customer	Gerling	IT	formatiert	Performance	Mips
Customer	Geschaeftskunden	IT	Formular	Performance	Nanosekunden
Customer	Gesundheitswesen	IT	Formulare	Performance	Performance
Customer	Gold-Zack	IT	Formularen	Performance	Plattenkapazitaet
Customer	Grosskunden	IT	FORTRAN	Performance	Rank

Dim	Term	Dim	Term	Dim	Term
Customer	hoechst	IT	Foto	Performance	Rechenleistung
Customer	ICL	IT	Fotos	Performance	schnell
Customer	Kaeufer	IT	Fragebogen	Performance	schnelle
Customer	Karstadt	IT	Frame	Performance	schnellen
Customer	KHK	IT	Framework	Performance	schnellere
Customer	Klientel	IT	Freigabe	Performance	Sekunden
Customer	Krankenhaeuser	IT	freigegeben	Performance	Speicherkapazitaet
Customer	Krupp	IT	freigegebenen	Performance	Taktfrequenz
Customer	Kunde	IT	Fremdsoftware	Performance	Taktrate
Customer	Kundschaft	IT	Frequenzen	Performance	UEbertragungsgeschwindigkeit
Customer	Lufthansa	IT	Fuehler	Performance	Uebertragungsrate
Customer	Mainframer	IT	Funk	Performance	UEbertragungsraten
Customer	Mensch	IT	Funktionalitaet	Performance	Wartezeiten
Customer	Messegesellschaft	IT	Funktionspruefung	Performance	Zugriffszeit
Customer	Nutzer	IT	Funktionstasten	Performance	Zuverlaessigkeit
Customer	Patienten	IT	Funktionsumfang	Performance	Zykluszeit
Customer	PC-Anwender	IT	Funktionsweise	Profession	Abiturienten
Customer	Post	IT	Fuzzy	Profession	Absolventen
Customer	RBG	IT	Gate-Arrays	Profession	absolviert
Customer	RTL	IT	Gates	Profession	Abteilungsleiter
Customer	RWE	IT	Gateway	Profession	Administrator
Customer	Schott	IT	Gateways	Profession	Administratoren
Customer	Shell	IT	gedruckt	Profession	Akademiker
Customer	Sparkassen	IT	gedruckte	Profession	Analyst
Customer	Surfer	IT	gedruckten	Profession	Analysten
Customer	Thyssen	IT	Gehaeuse	Profession	Anforderungsprofil
Customer	User	IT	Generator	Profession	Arbeiter
Customer	users	IT	Generatoren	Profession	Arbeitsamt
Customer	VKB	IT	Geraet	Profession	Arzt
Customer	Wall	IT	Geraete	Profession	Assistent
Customer	Warner	IT	Geraeten	Profession	Assistenten
Economy	Abnehmer	IT	Geraetes	Profession	Aufsichtsratsvorsitzender
Economy	Absatzchancen	IT	Gesamtkonzept	Profession	ausbilden
Economy	AG	IT	Gesamtpaket	Profession	Ausbilder
Economy	Airlines	IT	Gesamtsystem	Profession	Ausbildung
Economy	Aktie	IT	Gesamtsystems	Profession	ausgebildet
Economy	Aktien	IT	gespeichert	Profession	ausgebildete
Economy	Aktienanlagen	IT	gespeicherte	Profession	ausgebildeten
Economy	Aktiengesellschaft	IT	gespeicherten	Profession	auszubilden
Economy	Aktienkurs	IT	gesteuert	Profession	Beamten
Economy	Aktionaere	IT	gesteuerte	Profession	Berater
Economy	anbieten	IT	gesteuerten	Profession	Beratern
Economy	Anbieter	IT	Glasfaser	Profession	Beraters
Economy	Anbietern	IT	Glasfaserkabel	Profession	Beratung
Economy	Anbieters	IT	Global	Profession	Beratungsfirmen
Economy	Angebot	IT	Grafik	Profession	Beratungsgesellschaft
Economy	Angebote	IT	Grafiken	Profession	Beratungsleistung
Economy	angeboten	IT	grafisch	Profession	Beratungsleistungen
Economy	angebotene	IT	grafische	Profession	Beratungsunternehmen
Economy	angebotenen	IT	grafischen	Profession	Bereichsleiter
Economy	Angebotes	IT	grafischer	Profession	Beruf
Economy	Angebots	IT	grafisches	Profession	Berufe
Economy	Anleger	IT	Graphics	Profession	beruflich
Economy	Anwenderunternehmen	IT	graphische	Profession	berufliche
Economy	anzubieten	IT	graphischen	Profession	beruflichen
Economy	Application-Service-Provider	IT	Gross-EDV	Profession	Berufsausbildung
Economy	Arbeit	IT	Grossrechner	Profession	Berufsbild
Economy	Arbeiten	IT	Grossrechnern	Profession	Berufsbilder
Economy	Arbeitgeber	IT	Groupware	Profession	Berufserfahrung
Economy	Arbeitgebern	IT	Grundausstattung	Profession	Berufsgruppe
Economy	Arbeitgebers	IT	Grundfunktionen	Profession	Berufsgruppen
Economy	Arbeitslosen	IT	Grundkonfiguration	Profession	Berufsleben
Economy	Arbeitslosigkeit	IT	Grundsoftware	Profession	Berufspraxis
Economy	Arbeitsmarkt	IT	Halbleiter	Profession	Betriebswirt
Economy	Arbeitsteilung	IT	Handbuch	Profession	Betriebswirte
Economy	ASPs	IT	Handhelds	Profession	Bildung
Economy	Aufschwung	IT	Handy	Profession	CEO
Economy	Auftraggeber	IT	Handys	Profession	Chefredakteur
Economy	Auftraggebers	IT	Hardware	Profession	Chief
Economy	Auftragnehmer	IT	hardware-	Profession	CIO
Economy	Aufwand	IT	Hardwarebereich	Profession	CIOs
Economy	Aufwands	IT	Hardwareseitig	Profession	Consultant
Economy	Automobilindustrie	IT	Hauptrechner	Profession	Consultants
Economy	Bank	IT	Hauptspeicher	Profession	COO
Economy	Banken	IT	Hauptspeichers	Profession	Datenbankadministrator
Economy	Banking	IT	Heimcomputer	Profession	Datenschutzbeauftragte
Economy	bar	IT	Herstellerangaben	Profession	Datenschutzbeauftragten
Economy	Bargeld	IT	heterogenen	Profession	Datenschutzbeauftragter
Economy	Bedarf	IT	HIPO	Profession	Datenverarbeiter
Economy	Bedarfs	IT	hochintegrierte	Profession	Designer
Economy	Beratungshaus	IT	Homepage	Profession	Dienstleister
Economy	beschaeftigen	IT	Host-Rechner	Profession	Dipl
Economy	beschaeftigt	IT	Hosts	Profession	Diplom
Economy	Beschaeftigte	IT	Hotspots	Profession	Direktor
Economy	Beschaeftigten	IT	Hub	Profession	Dozent
Economy	Beschaeftigung	IT	Hubs	Profession	Dozenten
Economy	Betrag	IT	IBM-Produkte	Profession	Dr
Economy	Bezahlen	IT	IBM-Rechner	Profession	DSB
Economy	bezahlt	IT	IBM-Software	Profession	DV-Ausbildung
Economy	bezahlte	IT	IBM-Systeme	Profession	DV-Chef
Economy	Bezahlung	IT	IC	Profession	DV-Leiter
Economy	billig	IT	ICs	Profession	DV-Leute
Economy	billige	IT	IKS	Profession	DV-Manager

Dim	Term	Dim	Term	Dim	Term
Economy	billigen	IT	Imaging	Profession	DV-Mitarbeiter
Economy	billiger	IT	implementation	Profession	DV-Profis
Economy	billigere	IT	Implementieren	Profession	DV-Spezialisten
Economy	Billigste	IT	implementiert	Profession	DV-Verantwortlichen
Economy	Boerse	IT	implementierten	Profession	EDV-Ausbildung
Economy	Boersengang	IT	Implementierung	Profession	EDV-Beratung
Economy	Branche	IT	IMS	Profession	EDV-Chef
Economy	Branchen	IT	INA	Profession	EDV-Leiter
Economy	Branchenkenner	IT	Inbetriebnahme	Profession	EDV-Spezialisten
Economy	branchenspezifische	IT	Index	Profession	Einkaeufer
Economy	branchenspezifischen	IT	Industrieroboter	Profession	Entwickler
Economy	Business	IT	Informatik	Profession	Entwicklern
Economy	Co	IT	Information	Profession	Experte
Economy	Commercial	IT	Information-Highway	Profession	Fachleute
Economy	Company	IT	Informationen	Profession	Fachleuten
Economy	Computerbranche	IT	Informations	Profession	Fachmann
Economy	Computerhersteller	IT	Informationsflusses	Profession	Firmensprecher
Economy	Computerherstellers	IT	Informationsmanagement	Profession	Fortbildung
Economy	Computerindustrie	IT	Informationssystem	Profession	Geschaeftsfuehrer
Economy	Computermarkt	IT	Informationssysteme	Profession	Geschaeftsfuehrung
Economy	Corp	IT	Informationssystemen	Profession	Geschult
Economy	Corporate	IT	Informationssystems	Profession	Grundkenntnisse
Economy	Customer	IT	Informationstechnik	Profession	Grundwissen
Economy	Deal	IT	Informationstechnologie	Profession	Gruppenleiter
Economy	Dienstleistung	IT	Informationstechnologien	Profession	Hacker
Economy	Dienstleistungen	IT	Informationsverarbeitung	Profession	Handwerker
Economy	Dienstleistungsangebot	IT	Informatique	Profession	Headhunter
Economy	Dienstleistungsunternehmen	IT	Infrastructure	Profession	hochqualifizierte
Economy	Dotcoms	IT	Infrastruktur	Profession	hochqualifizierten
Economy	DV-Bereich	IT	inkompatibel	Profession	Hochschulabsolventen
Economy	DV-Branche	IT	Input	Profession	Hochschullehrer
Economy	DV-Industrie	IT	Installation	Profession	Informatiker
Economy	DV-Unternehmen	IT	Installationen	Profession	Informatikern
Economy	Economy	IT	Installieren	Profession	Ing
Economy	effektiv	IT	Installiert	Profession	Ingenieur
Economy	effektive	IT	installierte	Profession	Ingenieure
Economy	effektiven	IT	installierten	Profession	Ingenieuren
Economy	effektiver	IT	installierter	Profession	Insolvenzverwalter
Economy	effizient	IT	Instruktionen	Profession	IT-Chefs
Economy	effiziente	IT	Integrated	Profession	IT-Dienstleister
Economy	effizienten	IT	Integration	Profession	IT-Leiter
Economy	effizienter	IT	integrieren	Profession	IT-Manager
Economy	effizientere	IT	integriert	Profession	IT-Profis
Economy	Effizienz	IT	integrierte	Profession	IT-Spezialisten
Economy	Einkommen	IT	integriertem	Profession	IT-Verantwortlichen
Economy	Einzelhandel	IT	integrierten	Profession	Journalist
Economy	Einzelhandels	IT	Integrierter	Profession	Journalistin
Economy	Elektroindustrie	IT	integriertes	Profession	Karriere
Economy	Elektrokonzern	IT	Interface	Profession	Kfm
Economy	Endverbraucher	IT	Interfaces	Profession	Lehrer
Economy	Energie	IT	Internet	Profession	Leiter
Economy	Energieversorgung	IT	Internet-	Profession	Lieferant
Economy	erworben	IT	Internet-Telefonie	Profession	Marktbeobachter
Economy	erworbenen	IT	Internet-Zugang	Profession	Netzwerker
Economy	Export	IT	Internetworking	Profession	Officer
Economy	exportieren	IT	Interoperabilitaet	Profession	Organisatoren
Economy	exportiert	IT	Intranet	Profession	Personalberater
Economy	Fachhandel	IT	Intranets	Profession	Praesident
Economy	Fertigungsindustrie	IT	ISDN	Profession	Prof
Economy	Festpreis	IT	IT	Profession	Professor
Economy	Finanzdienstleister	IT	IT-	Profession	Programmierer
Economy	Firma	IT	IT-Branche	Profession	Programmierern
Economy	Firmen	IT	IT-Infrastruktur	Profession	Projektleiter
Economy	Fluggesellschaften	IT	IT-Sicherheit	Profession	Qualifikation
Economy	Fluktuation	IT	IT-Systeme	Profession	Qualifikationen
Economy	Foerderung	IT	Jahr-2000-Problem	Profession	qualifizierte
Economy	Fusion	IT	Kabel	Profession	qualifizierten
Economy	Garantie	IT	Kanaele	Profession	Qualifizierung
Economy	garantieren	IT	Karte	Profession	Rechtsanwalt
Economy	garantiert	IT	Karten	Profession	Referenten
Economy	gebot	IT	Kartenleser	Profession	Sachbearbeiter
Economy	geboten	IT	Kassette	Profession	Schulung
Economy	gebotenen	IT	Kassetten	Profession	Schulungen
Economy	gekauft	IT	KB	Profession	Selbstaendige
Economy	gekaufte	IT	KBit	Profession	Seminar
Economy	gekauften	IT	Kernel	Profession	Software-Entwickler
Economy	Geld	IT	Kernspeicher	Profession	Softwarehaus
Economy	Gelder	IT	Key	Profession	Spezialist
Economy	Geldgeber	IT	KI	Profession	Spezialisten
Economy	Geldinstitute	IT	Kilobyte	Profession	Steuerberater
Economy	Geldinstituten	IT	Kit	Profession	Systemanalytiker
Economy	Gesamtwert	IT	Klartext	Profession	Systemhaus
Economy	Geschaeft	IT	Klasse	Profession	Techniker
Economy	Geschaefte	IT	Klassen	Profession	Topmanager
Economy	Geschaeften	IT	Kleincomputer	Profession	Unternehmensberater
Economy	Geschaefts	IT	Kleinrechner	Profession	Unternehmensberatung
Economy	Geschaeftsbereich	IT	Klimaanlage	Profession	Unternehmensfuehrung
Economy	Geschaeftsjahr	IT	Knoten	Profession	Unternehmensleitung
Economy	Geschaeftsjahres	IT	Knowledge-Management	Profession	Unternehmer
Economy	Geschaeftsleitung	IT	Komfort	Profession	Verfasser
Economy	Geschaeftsprozesse	IT	Kommunikation	Profession	Verkaeufer
Economy	Geschaeftsprozessen	IT	Kommunikations-	Profession	Verleiher
Economy	Geschaeftsstelle	IT	Kommunikationsanalyse	Profession	Vertriebsbeauftragten

Dim	Term	Dim	Term	Dim	Term
Economy	Geschaeftsstellen	IT	Kommunikationssysteme	Profession	Vertriebsleiter
Economy	Gesellschaften	IT	Kommunikationstechnik	Profession	Vice-President
Economy	Gesellschafter	IT	kommunizieren	Profession	Vorgesetzten
Economy	globalen	IT	kompatibel	Profession	Vorstand
Economy	Grossunternehmen	IT	Kompatibilitaet	Profession	Vorstandsmitglied
Economy	Group	IT	Komplettloesung	Profession	Vorstandsvorsitzender
Economy	Grundpreis	IT	komplexe	Profession	Weiterbildung
Economy	Haftung	IT	Komplexitaet	ProgLanguage	4GL
Economy	Handel	IT	Komponente	ProgLanguage	ADA
Economy	Hardwarehersteller	IT	Komponenten	ProgLanguage	APL
Economy	Hermes	IT	Konfiguration	ProgLanguage	C
Economy	Hersteller	IT	Konfigurationen	ProgLanguage	C+
Economy	Herstellern	IT	konfigurieren	ProgLanguage	EJB
Economy	Herstellers	IT	Konzept	ProgLanguage	Forth
Economy	Herstellerseite	IT	Konzepte	ProgLanguage	J2EE
Economy	Import	IT	Konzeption	ProgLanguage	Java
Economy	importiert	IT	konzipiert	ProgLanguage	Java-Applets
Economy	Inc	IT	Kopien	ProgLanguage	Javabeans
Economy	Industrie	IT	Koppelung	ProgLanguage	JVM
Economy	Industriebetrieben	IT	Kopplung	ProgLanguage	LISP
Economy	Industrien	IT	Kundendaten	ProgLanguage	Mantis
Economy	Industrienationen	IT	LAN	ProgLanguage	Pascal
Economy	Industries	IT	LAN-	ProgLanguage	Perl
Economy	Industrieunternehmen	IT	Language	ProgLanguage	PL/1
Economy	Industriezweig	IT	LANs	ProgLanguage	Prolog
Economy	Industriezweige	IT	Laptop	ProgLanguage	RPG
Economy	Industriezweigen	IT	Laptops	ProgLanguage	Smalltalk
Economy	Industry	IT	Laserdrucker	ProgLanguage	SQL
Economy	Insolvenz	IT	Laufwerk	ProgLanguage	SQL/DS
Economy	Interessenten	IT	Laufwerke	Science	Akademie
Economy	Internet-Firmen	IT	Layer	Science	akademischen
Economy	Internet-Service-Provider	IT	LB	Science	Basiswissen
Economy	Investitionszulage	IT	Learning	Science	Betriebswirtschaftslehre
Economy	Investoren	IT	Leitungen	Science	Cambridge
Economy	ISP	IT	Link	Science	Fachhochschule
Economy	ISPs	IT	Liste	Science	Fachhochschulen
Economy	IT-Dienstleistungen	IT	Listen	Science	Fachkenntnisse
Economy	IT-Industrie	IT	Locher	Science	Fachliteratur
Economy	IT-Unternehmen	IT	Lochkarte	Science	Fachrichtung
Economy	Job	IT	Lochkarten	Science	Fachrichtungen
Economy	Jobs	IT	Lochstreifen	Science	Fachtagung
Economy	Joint	IT	Loesung	Science	Fachwelt
Economy	Joint-venture	IT	Loesungen	Science	Fachwissen
Economy	Kauf	IT	Logic	Science	FORMEL
Economy	kaufen	IT	Logik	Science	Formeln
Economy	Kaufpreis	IT	LVN	Science	Forscher
Economy	KG	IT	M-Commerce	Science	Forschung
Economy	kommerzielle	IT	Machine	Science	Forschungen
Economy	Konjunktur	IT	Machines	Science	Forschungseinrichtungen
Economy	Konkurrent	IT	Magnetbaender	Science	Forschungsgruppe
Economy	Konkurrenten	IT	Magnetband	Science	Forschungsinstitut
Economy	Konkurrenz	IT	Magnetbandkassette	Science	Forschungsinstitute
Economy	Konsortium	IT	Magnetkonto	Science	Forschungsprojekt
Economy	Kooperation	IT	Magnetplatte	Science	Forschungsprojekte
Economy	Kosten	IT	Magnetplatten	Science	Forum
Economy	kostet	IT	Mail	Science	Grundlagenforschung
Economy	Krise	IT	Mailbox	Science	Hochschule
Economy	Kunden	IT	Mainframe	Science	Hochschulen
Economy	Kurse	IT	Mainframe-	Science	IDC
Economy	Leasing	IT	Mainframes	Science	Institut
Economy	Leasing-Gesellschaft	IT	Markup	Science	Institute
Economy	Leasing-Gesellschaften	IT	Maschine	Science	Instituten
Economy	Leasinggeber	IT	maschinellen	Science	Institution
Economy	Leasingnehmer	IT	Masken	Science	Institutionen
Economy	Lieferanten	IT	Massenspeicher	Science	Instituts
Economy	liefern	IT	Matrixdrucker	Science	Knowledge
Economy	Lieferumfang	IT	Maus	Science	Konferenz
Economy	Lizenzen	IT	MB	Science	Lehre
Economy	Logistik	IT	MDT	Science	Lehrstuhl
Economy	LTD	IT	MDT-Anlagen	Science	Literatur
Economy	Maerkte	IT	MDT-Computer	Science	Marktforscher
Economy	Maerkten	IT	Medien	Science	Mathematik
Economy	Marke	IT	Medium	Science	Methode
Economy	Market	IT	Memory	Science	Pruefen
Economy	Markt	IT	Menues	Science	Pruefung
Economy	Marktanteil	IT	Message	Science	Schule
Economy	Marktanteile	IT	Messaging	Science	Schulen
Economy	Marktes	IT	Metadaten	Science	Seminare
Economy	Marktfuehrer	IT	Methoden	Science	Seminaren
Economy	Marktfuehrers	IT	MIB	Science	Studenten
Economy	Marktplaetze	IT	MIBs	Science	Studie
Economy	Marktplaetzen	IT	Michelangelo	Science	Studien
Economy	Marktplatz	IT	Micro	Science	Studierenden
Economy	Marktsegment	IT	Microsystems	Science	Studium
Economy	Maschinenbau	IT	Middleware	Science	technisch-wissenschaftlichen
Economy	Massenmarkt	IT	Migration	Science	Theoretisch
Economy	mbH	IT	Mikro	Science	Theorie
Economy	MDT-Hersteller	IT	Mikrocode	Science	Training
Economy	Merger	IT	Mikrocomputer	Science	Uni
Economy	Messe	IT	Mikrocomputern	Science	Universitaet
Economy	Mieter	IT	Mikrocomputers	Science	Universitaeten
Economy	Mittelbetriebe	IT	Mikroelektronik	Science	University
Economy	mittelstaendische	IT	Mikrofiches	Science	Unterricht

Dim	Term	Dim	Term	Dim	Term
Economy	mittelstaendischen	IT	Mikrofilm	Science	untersucht
Economy	Mittelstaendler	IT	Mikroprozessor	Science	Untersuchungen
Economy	Monatsmiete	IT	Mikroprozessoren	Science	Vorgehensweise
Economy	Nachfrage	IT	Mikros	Science	Vortraege
Economy	Nasdaq	IT	Mini	Science	Vortrag
Economy	Netzbetreiber	IT	Minicomputer	Science	Wirtschaftsinformatik
Economy	Nutzen	IT	Minicomputern	Science	Wissens
Economy	Nutzung	IT	Minis	Science	Wissensbasierte
Economy	operativen	IT	MIS	Science	Wissensbasis
Economy	Outsourcer	IT	Mittelstand	Science	Wissenschaft
Economy	Outsourcing	IT	mobiler	Science	Wissenschaftler
Economy	Partner	IT	Mobilfunk	Science	wissenschaftliche
Economy	Partnerschaft	IT	Mobiltelefon	Science	wissenschaftlichen
Economy	Partnerschaften	IT	Modelle	SocialFramework	Amt
Economy	PC-Geschaeft	IT	Modellen	SocialFramework	Amtes
Economy	PC-Markt	IT	Modem	SocialFramework	Aufsichtsbehoerde
Economy	PCMer	IT	Modems	SocialFramework	Aufsichtsbehoerden
Economy	PCMs	IT	Modul	SocialFramework	AWV
Economy	Plc	IT	modular	SocialFramework	BDSB
Economy	Portfolio	IT	Module	SocialFramework	BDSG
Economy	Preise	IT	Moduln	SocialFramework	Behoerde
Economy	Preisen	IT	Monitor	SocialFramework	Behoerden
Economy	President	IT	Multimedia	SocialFramework	Bestimmung
Economy	Produkten	IT	Multimedia-	SocialFramework	Bestimmungen
Economy	PTT	IT	Multiplexer	SocialFramework	BfA
Economy	PTTs	IT	Multiprogramming	SocialFramework	BGB
Economy	Ressourcen	IT	Nachdokumentation	SocialFramework	BMFT
Economy	Rezession	IT	Nachricht	SocialFramework	Bund
Economy	Schaeden	IT	Nachrichtentechnik	SocialFramework	Bundes
Economy	Service	IT	NAS	SocialFramework	Bundesamt
Economy	Services	IT	NC	SocialFramework	Bundesanstalt
Economy	Software-Anbieter	IT	NCs	SocialFramework	Bundesdatenschutzgesetz
Economy	Software-Industrie	IT	Nebenrechner	SocialFramework	bundesdeutschen
Economy	Software-Unternehmen	IT	Nebenstellenanlagen	SocialFramework	Bundesministerium
Economy	Softwareanbieter	IT	net	SocialFramework	Bundesministeriums
Economy	Softwarehaeuser	IT	Network	SocialFramework	Bundesregierung
Economy	Softwarehaeusern	IT	Networking	SocialFramework	Bundestag
Economy	Softwarehauses	IT	Networks	SocialFramework	Bundestages
Economy	Softwaremarkt	IT	Netz	SocialFramework	Bundesverband
Economy	Softwareschmiede	IT	Netz-	SocialFramework	Bundesverbandes
Economy	Softwareunternehmen	IT	Netze	SocialFramework	Datenschutz
Economy	sparen	IT	Netzen	SocialFramework	Datenschutzes
Economy	spart	IT	Netzes	SocialFramework	Datenschutzgesetz
Economy	Standort	IT	Netzwerk	SocialFramework	DOJ
Economy	Standorte	IT	Netzwerk-	SocialFramework	e.V
Economy	Startup	IT	Netzwerk-Management	SocialFramework	EG
Economy	Startups	IT	Netzwerke	SocialFramework	EG-Kommission
Economy	Statistik	IT	Netzwerken	SocialFramework	EU
Economy	Statistiken	IT	Neuronale	SocialFramework	FTC
Economy	Stiftung	IT	neuronalen	SocialFramework	GDD
Economy	Systemhaeuser	IT	Neuronaler	SocialFramework	genehmigt
Economy	Teilnahmegebuehr	IT	NFS	SocialFramework	Genehmigung
Economy	teuer	IT	Node	SocialFramework	Gericht
Economy	Transit	IT	Notebook	SocialFramework	Gerichte
Economy	uebernahm	IT	Notebooks	SocialFramework	Gerichten
Economy	UEbernahme	IT	Oberflaeche	SocialFramework	Gesetz
Economy	UEbernahmen	IT	Oberflaechen	SocialFramework	Gesetze
Economy	uebernehmen	IT	Object	SocialFramework	Gesetzes
Economy	uebernimmt	IT	Object-class	SocialFramework	Gesetzgeber
Economy	uebernommen	IT	Object-instance	SocialFramework	Gesetzgebung
Economy	Unbundling	IT	Objects	SocialFramework	gesetzlich
Economy	Unternehmen	IT	Objekt	SocialFramework	gesetzliche
Economy	unternehmensweite	IT	Objekten	SocialFramework	gesetzlichen
Economy	unternehmensweiten	IT	objektorientierte	SocialFramework	Gewerkschaft
Economy	Unternehmungen	IT	objektorientierten	SocialFramework	Gewerkschaften
Economy	US-Markt	IT	objektorientierter	SocialFramework	Handelskammer
Economy	US-Unternehmen	IT	objektorientiertes	SocialFramework	Handelskammern
Economy	VCs	IT	Objektorientierung	SocialFramework	IDA
Economy	verdienen	IT	OCR	SocialFramework	IG-EDV
Economy	Verkauf	IT	Offenheit	SocialFramework	Informationsgesellschaft
Economy	verkaufen	IT	offline	SocialFramework	Innenministerium
Economy	verkauft	IT	Olap	SocialFramework	Jessi
Economy	verkaufte	IT	OLE	SocialFramework	Klaeger
Economy	Verlag	IT	Online	SocialFramework	Klaegerin
Economy	Vermieter	IT	Online-	SocialFramework	Kommunen
Economy	Versicherer	IT	Online-Dienst	SocialFramework	Liberalisierung
Economy	Versicherung	IT	Online-Dienste	SocialFramework	Medizin
Economy	Versicherungen	IT	Online-Programmierung	SocialFramework	Ministerium
Economy	Wachstum	IT	Online-Shops	SocialFramework	MITI
Economy	Wachstumsraten	IT	Online-Systemen	SocialFramework	Natur
Economy	Ware	IT	Online-Verarbeitung	SocialFramework	NI
Economy	Weltmarkt	IT	OOP	SocialFramework	NSA
Economy	wert	IT	Open-Source-Software	SocialFramework	OEffentlichkeit
Economy	Wettbewerb	IT	Operating	SocialFramework	Organisationen
Economy	Wettbewerber	IT	Operationen	SocialFramework	Paragraph
Economy	Wettbewerbern	IT	Operations	SocialFramework	Paragraphen
Economy	Wettbewerbsfaehigkeit	IT	Operator	SocialFramework	Parteien
Economy	Wirtschaft	IT	Operatoren	SocialFramework	Politik
Economy	wirtschaftlich	IT	optimieren	SocialFramework	Rahmenbedingungen
Economy	wirtschaftlichen	IT	optimiert	SocialFramework	Regelung
Economy	wirtschaftlicher	IT	Optimierung	SocialFramework	Regelungen
Economy	Wirtschaftlichkeit	IT	Orgatechnik	SocialFramework	Regierung
Economy	zahlt	IT	Output	SocialFramework	Richtlinie

Dim	Term	Dim	Term	Dim	Term
Economy	Zielgruppe	IT	Papier	SocialFramework	Richtlinien
Economy	Zinsen	IT	Paradox	SocialFramework	Soziale
Economy	Zulieferer	IT	Parallelrechner	SocialFramework	sozialen
Economy	Zulieferern	IT	Parameter	SocialFramework	Staat
Economy	Zuwachsraten	IT	Partition	SocialFramework	staatlich
Event	Aussteller	IT	Partitions	SocialFramework	staatliche
Event	Ausstellern	IT	Passwort	SocialFramework	staatlichen
Event	Ausstellung	IT	Patterns	SocialFramework	TUEV
Event	Ausstellungen	IT	PC	SocialFramework	Umwelt
Event	CeBIT	IT	PC-	SocialFramework	Urteil
Event	CeBIT-Nord	IT	PC-Hersteller	SocialFramework	Urteile
Event	Comdex	IT	PCs	SocialFramework	VDRZ
Event	COMM-PRIX	IT	PDA	SocialFramework	Verband
Event	Euroforum	IT	PERFORM	SocialFramework	Verein
Event	Expo	IT	Peripherie	SocialFramework	Vorschriften
Event	Fachbesucher	IT	Peripheriegeraete	SocialFramework	ZAV
Event	Fachmesse	IT	Personalcomputer	SocialFramework	ZVEI
Event	Fachmessen	IT	Personalcomputern	Vendor	3Com
Event	Hannover-Messe	IT	Phasen	Vendor	Abb
Event	Infoworld	IT	Pixel	Vendor	ACE
Event	Kongress	IT	PKI	Vendor	Acer
Event	Messe-	IT	Planungssprachen	Vendor	Adler
Event	MMG	IT	Platform	Vendor	Adobe
Event	Symposium	IT	Platte	Vendor	ADR
Event	Tagung	IT	Platten	Vendor	ADV/ORGA
Event	Veranstalter	IT	Plattenspeicher	Vendor	AEG
Event	veranstaltet	IT	Plattform	Vendor	AEG-Telefunken
Event	Veranstaltung	IT	Plattformen	Vendor	Alcatel
Event	Veranstaltungen	IT	Player	Vendor	Altavista
Geography	Aachen	IT	PLM	Vendor	Amazon
Geography	Aachener	IT	Plotter	Vendor	AMD
Geography	Afrika	IT	Pocket	Vendor	Amdahl
Geography	America	IT	Points	Vendor	Ameritech
Geography	American	IT	Port	Vendor	Anker
Geography	Amerika	IT	Portabilitaet	Vendor	AOL
Geography	Amerikaner	IT	Portal	Vendor	Apex
Geography	Amerikanern	IT	Portale	Vendor	Apollo
Geography	amerikanische	IT	Portals	Vendor	Apple
Geography	amerikanischen	IT	portieren	Vendor	Apples
Geography	amerikanischer	IT	portiert	Vendor	Appware
Geography	Amerikanisches	IT	Portierung	Vendor	Arcor
Geography	Amsterdam	IT	Portlets	Vendor	Ariba
Geography	Angeles	IT	Ports	Vendor	Ascend
Geography	arabischen	IT	PPS	Vendor	Ashton-Tate
Geography	ARMONK	IT	PPS-System	Vendor	AT
Geography	Armonker	IT	Presentation	Vendor	Atari
Geography	Asien	IT	Processing	Vendor	Atlantic
Geography	Augsburg	IT	Program	Vendor	Baan
Geography	Augsburger	IT	Programm	Vendor	Banyan
Geography	Ausland	IT	Programmdokumentation	Vendor	BASF
Geography	Australien	IT	Programme	Vendor	Bay
Geography	Basel	IT	Programmen	Vendor	Bayer
Geography	bayerische	IT	Programmentwicklung	Vendor	Bea
Geography	Bayerischen	IT	Programmes	Vendor	Bell
Geography	Bayern	IT	Programmieren	Vendor	Bertelsmann
Geography	Belgien	IT	Programmiersprache	Vendor	BMW
Geography	belgische	IT	Programmiersprachen	Vendor	Borland
Geography	belgischen	IT	Programmierung	Vendor	Bosch
Geography	BERLIN	IT	Programmpaket	Vendor	BT
Geography	Berliner	IT	Programms	Vendor	Bull
Geography	Bern	IT	Programmstruktur	Vendor	Bundespost
Geography	Berner	IT	Programmsystem	Vendor	Burroughs
Geography	BIELEFELD	IT	Protocol	Vendor	CA
Geography	Bochum	IT	Protokoll	Vendor	Cabletron
Geography	BOEBLINGEN	IT	Protokolle	Vendor	Calcomp
Geography	BONN	IT	Prototyp	Vendor	CDC
Geography	Bonner	IT	prototypen	Vendor	Centronics
Geography	BOSTON	IT	Prototyping	Vendor	Ceyoniq
Geography	Brandenburg	IT	Provider	Vendor	Chipcom
Geography	Brasilien	IT	Prozedur	Vendor	Ciena
Geography	Braunschweig	IT	Prozeduren	Vendor	Cisco
Geography	Bremen	IT	Prozess	Vendor	CMB
Geography	Bremer	IT	Prozessor	Vendor	Cognos
Geography	Briten	IT	Prozessoren	Vendor	Commodore
Geography	britische	IT	Prozessrechner	Vendor	Compagnie
Geography	britischen	IT	Publishing	Vendor	Compaq
Geography	British	IT	Qualitaet	Vendor	Compaqs
Geography	Bruessel	IT	Qualitaetssicherung	Vendor	Comparex
Geography	bundesdeutsche	IT	Quellcode	Vendor	Compunet
Geography	Bundesgebiet	IT	Quellenauswahl	Vendor	Compuserve
Geography	Bundeslaendern	IT	Radio	Vendor	Computerland
Geography	Bundesland	IT	RAM	Vendor	Convex
Geography	Bundesrepublik	IT	RDBMS	Vendor	Corel
Geography	Cairo	IT	Re-Engineering	Vendor	Cray
Geography	California	IT	Realisierung	Vendor	CTM
Geography	Canada	IT	Realtime	Vendor	Cullinet
Geography	Canadian	IT	Rechenzentren	Vendor	Cunningham
Geography	Chemnitz	IT	Rechenzentrum	Vendor	Cyrix
Geography	Chicago	IT	Rechenzentrums	Vendor	DatagraphiX
Geography	China	IT	rechnen	Vendor	Datapoint
Geography	DALLAS	IT	Rechner	Vendor	Datasaab
Geography	Darmstadt	IT	Rechnern	Vendor	Datev
Geography	DDR	IT	Rechners	Vendor	DBP

Dim	Term	Dim	Term	Dim	Term
Geography	Deutsch	IT	Rechnerverbund	Vendor	Debis
Geography	deutsche	IT	rechnet	Vendor	DEC
Geography	deutschen	IT	Rechnungen	Vendor	DECnet
Geography	deutscher	IT	Recovery	Vendor	DECs
Geography	deutsches	IT	Regel	Vendor	Dell
Geography	Deutschland	IT	Regeln	Vendor	DeTeMobil
Geography	Deutschlands	IT	Relational	Vendor	Diebold
Geography	deutschsprachigen	IT	Relationale	Vendor	Digital
Geography	Dortmund	IT	relationalen	Vendor	Digitals
Geography	Dortmunder	IT	Relay	Vendor	Ditec
Geography	Dresdner	IT	Release	Vendor	Documentum
Geography	Duesseldorf	IT	Remote	Vendor	EDS
Geography	Duesseldorfer	IT	Repository	Vendor	Elbit
Geography	Duisburg	IT	Request	Vendor	EMC
Geography	Eching	IT	Return	Vendor	Epson
Geography	Eiffel	IT	RFC	Vendor	Ericsson
Geography	Elsaesser	IT	RISC	Vendor	Escom
Geography	Emea	IT	Roboter	Vendor	Everex
Geography	England	IT	Robotern	Vendor	Facit
Geography	Englisch	IT	ROM	Vendor	Freenet
Geography	englische	IT	Router	Vendor	FSC
Geography	englischen	IT	Routine	Vendor	Fujitsu
Geography	englischer	IT	Routing	Vendor	Fujitsu-Siemens
Geography	Erde	IT	RZ-Betrieb	Vendor	Gates-Company
Geography	Eschborn	IT	SAA	Vendor	GE
Geography	Essen	IT	SANs	Vendor	Gemini
Geography	ESSLINGEN	IT	Satelliten	Vendor	GMO
Geography	Ettlingen	IT	Scanner	Vendor	Google
Geography	Europa	IT	Schaltungen	Vendor	Handspring
Geography	Europaeische	IT	Schicht	Vendor	Harris
Geography	europaeischen	IT	Schichten	Vendor	Heiler
Geography	Europas	IT	Schluessel	Vendor	Hewlett
Geography	Europe	IT	Schnelldrucker	Vendor	Hewlett-Packard
Geography	European	IT	Schnittstelle	Vendor	Hitachi
Geography	Finnland	IT	Schnittstellen	Vendor	Honeywell
Geography	Florida	IT	Schreibmaschinen	Vendor	HP
Geography	France	IT	SCM	Vendor	HPs
Geography	Francisco	IT	SDLC	Vendor	Hyperion
Geography	Frankfurt	IT	Search	Vendor	I2
Geography	FRANKFURT/M	IT	Secure	Vendor	IBM
Geography	Frankfurt/Main	IT	Security	Vendor	IBM-
Geography	Frankfurter	IT	Sektoren	Vendor	IBM-Tochter
Geography	Frankreich	IT	Senden	Vendor	IBMs
Geography	franzoesische	IT	serielle	Vendor	IDV
Geography	franzoesischen	IT	Server	Vendor	Infineon
Geography	Franzosen	IT	Server-	Vendor	Inforex
Geography	Freiburg	IT	Servern	Vendor	Informix
Geography	Freising	IT	Servers	Vendor	Inmos
Geography	Friedrichshafen	IT	Service-	Vendor	Inprise
Geography	Genf	IT	Service-Provider	Vendor	Intel
Geography	German	IT	Service-Rechenzentren	Vendor	Intels
Geography	Germany	IT	Service-Rechenzentrum	Vendor	Interdata
Geography	GRASBRUNN	IT	Session	Vendor	Intergraph
Geography	Graz	IT	Set	Vendor	Interkom
Geography	Griechenland	IT	Sicherheit	Vendor	Intershop
Geography	Grossbritannien	IT	Sicherheitsluecken	Vendor	Intuit
Geography	Gummersbach	IT	Signatur	Vendor	IT-Anbieter
Geography	Haag	IT	Signaturen	Vendor	Itos
Geography	Halle	IT	Silicon	Vendor	ITT
Geography	Hamburg	IT	Simulation	Vendor	Ixos
Geography	Hamburger	IT	simulieren	Vendor	Kienzle
Geography	HANNOVER	IT	Site	Vendor	Kodak
Geography	Heidelberg	IT	Sites	Vendor	Kombinat
Geography	Heilbronn	IT	Skalierbare	Vendor	Kontron
Geography	Hessen	IT	Skalierbarkeit	Vendor	KPN
Geography	Hessischen	IT	Smartphones	Vendor	Legent
Geography	Holland	IT	SMDS	Vendor	Lexmark
Geography	Homburg	IT	SMS	Vendor	Logabax
Geography	Hongkong	IT	Software	Vendor	Lucent
Geography	Houston	IT	Software-	Vendor	Lycos
Geography	ICC	IT	Software-Engineering	Vendor	Mannesmann
Geography	Illinois	IT	Software-Entwicklung	Vendor	Matsushita
Geography	Indien	IT	Software-Paket	Vendor	MBB
Geography	indische	IT	Software-Qualitaetssicherung	Vendor	mbp
Geography	indischen	IT	Softwareentwicklung	Vendor	McAfee
Geography	Inland	IT	Softwarehersteller	Vendor	MCI
Geography	International	IT	Softwareloesungen	Vendor	MDS
Geography	Internationale	IT	Softwarepaket	Vendor	Memorex
Geography	internationalen	IT	Softwarepakete	Vendor	Mergard
Geography	internationaler	IT	Softwareprodukte	Vendor	Microsoft
Geography	Irland	IT	Sort	Vendor	Microsoft-
Geography	Ismaning	IT	Sortieren	Vendor	Microsofts
Geography	Italia	IT	Source	Vendor	Microstrategy
Geography	Italien	IT	Sourcecode	Vendor	Mitsubishi
Geography	Japan	IT	Speicher	Vendor	Mobilcom
Geography	Japaner	IT	Speichermedien	Vendor	Mostek
Geography	japanische	IT	speichern	Vendor	Motorola
Geography	japanischen	IT	Speichersysteme	Vendor	MS
Geography	Japans	IT	speichert	Vendor	MSN
Geography	Kalifornien	IT	Speicherung	Vendor	NAI
Geography	Kalifornier	IT	Spezifikation	Vendor	Napster
Geography	kalifornische	IT	Spezifikationen	Vendor	Navision

Dim	Term	Dim	Term	Dim	Term
Geography	kalifornischen	IT	Sprache	Vendor	NCR
Geography	Kanada	IT	Spracheingabe	Vendor	NEC
Geography	Karlsruhe	IT	Stabilitaet	Vendor	Netscape
Geography	Kiel	IT	Stammdaten	Vendor	Netscapes
Geography	Koelner	IT	Standard	Vendor	Next
Geography	Konstanz	IT	Standard-Software	Vendor	Nixdorf
Geography	Land	IT	standardisierte	Vendor	Nokia
Geography	Leipzig	IT	standardisierten	Vendor	Norsk
Geography	Linz	IT	Standardisierung	Vendor	Nortel
Geography	London	IT	Standardprogramme	Vendor	Novell
Geography	LUDWIGSHAFEN	IT	Standards	Vendor	Novells
Geography	MAINZ	IT	Standardsoftware	Vendor	Offerto
Geography	Mannheim	IT	Stapelverarbeitung	Vendor	Olivetti
Geography	Massachusetts	IT	steckerkompatiblen	Vendor	Olympia
Geography	MENLO	IT	Steuereinheit	Vendor	Oracle
Geography	Moskau	IT	Steuereinheiten	Vendor	Oracles
Geography	MUENCHEN	IT	Steuerung	Vendor	Orgatec
Geography	Muenchener	IT	Storage	Vendor	Otelo
Geography	Muenchner	IT	Stromversorgung	Vendor	Packard
Geography	NEU-ISENBURG	IT	Struktogramme	Vendor	Palm
Geography	Neuss	IT	Struktur	Vendor	Parcplace
Geography	Nippon	IT	Studio	Vendor	Parsytec
Geography	Nord	IT	Suchmaschine	Vendor	Paybox
Geography	Nordrhein-Westfalen	IT	Suchmaschinen	Vendor	PeopleSoft
Geography	Nuernberg	IT	Suite	Vendor	Perkin-Elmer
Geography	Nuernberger	IT	Supercomputer	Vendor	Philips
Geography	OEsterreich	IT	Superminis	Vendor	Pixelpark
Geography	oesterreichische	IT	Supply-Chain-Management	Vendor	Planethome
Geography	oesterreichischen	IT	Support	Vendor	Platinum
Geography	Ort	IT	surfen	Vendor	Progress
Geography	Osten	IT	Switch	Vendor	Prologue
Geography	Osteuropa	IT	Switches	Vendor	Qwest
Geography	Paderborn	IT	Syntax	Vendor	Realnames
Geography	Paderborner	IT	System	Vendor	Retix
Geography	Paris	IT	System-	Vendor	Robotics
Geography	Polen	IT	System-Management	Vendor	Robotron
Geography	Redmond	IT	System-Software	Vendor	RSL
Geography	Redmonder	IT	Systemanalyse	Vendor	SAG
Geography	Regensburg	IT	Systeme	Vendor	Samsung
Geography	Region	IT	Systemen	Vendor	SAP
Geography	Regionen	IT	Systementwicklung	Vendor	SAPs
Geography	Schleswig-Holstein	IT	Systemintegration	Vendor	SAS
Geography	Schweden	IT	Systemkomponenten	Vendor	SCO
Geography	Schweiz	IT	Systems	Vendor	Seagate
Geography	Schweizer	IT	Systemsoftware	Vendor	Seebeyond
Geography	schweizerische	IT	Tabellenkalkulation	Vendor	SEL
Geography	Schweizerischen	IT	Tablet	Vendor	Semiconductor
Geography	skandinavischen	IT	Taschenrechner	Vendor	Sequent
Geography	Staaten	IT	Task	Vendor	SGI
Geography	Stuttgart	IT	Tastatur	Vendor	Sharp
Geography	Stuttgarter	IT	TC	Vendor	Siebel
Geography	Suisse	IT	Technik	Vendor	Siemens
Geography	Sulzbach	IT	technische	Vendor	Siemens-Nixdorf
Geography	Sydney	IT	technischen	Vendor	Singer
Geography	Texaner	IT	Technologie	Vendor	Sirius
Geography	Texas	IT	Technologien	Vendor	SMC
Geography	TOKIO	IT	Technologies	Vendor	SNI
Geography	U.S	IT	Technology	Vendor	Softlab
Geography	US-	IT	Telearbeit	Vendor	Sony
Geography	US-amerikanische	IT	Telecom	Vendor	SPC
Geography	USA	IT	Telecommunications	Vendor	Sperry
Geography	VALLEY	IT	Telefax	Vendor	Sprint
Geography	VILLINGEN	IT	Telefon	Vendor	SSA
Geography	Walldorfer	IT	Telegraph	Vendor	Stac
Geography	WASHINGTON	IT	Telekommunikation	Vendor	STC
Geography	Welt	IT	Telephone	Vendor	Stratus
Geography	weltweit	IT	Teleport	Vendor	Sun
Geography	weltweite	IT	Teleports	Vendor	Suns
Geography	weltweiten	IT	Teletex	Vendor	Sunsoft
Geography	Westeuropa	IT	Telex	Vendor	Suse
Geography	westlichen	IT	Terminal	Vendor	Sybase
Geography	Wien	IT	Terminals	Vendor	Symantec
Geography	Wiener	IT	Test	Vendor	Synoptics
Geography	Wiesbaden	IT	Testdaten	Vendor	Systec
Geography	Wiesbadener	IT	testen	Vendor	Systor
Geography	WILHELMSHAVEN	IT	Testhilfen	Vendor	T-Mobile
Geography	World	IT	Tests	Vendor	T-Online
Geography	Worms	IT	Text	Vendor	Taligent
Geography	Wuppertal	IT	Textautomaten	Vendor	Tandem
Geography	YORK	IT	Textsystem	Vendor	Tandon
Geography	Yorker	IT	Textsysteme	Vendor	Taylorix
Geography	ZUERICH	IT	Textverarbeitung	Vendor	Tektronix
Institute	Accenture	IT	Tischrechner	Vendor	Telekom
Institute	Andersen	IT	TK-	Vendor	Tewidata
Institute	apa	IT	TK-Anlagen	Vendor	TI
Institute	BBN	IT	Token	Vendor	Tibco
Institute	Bitkom	IT	Tool	Vendor	Toshiba
Institute	Brancheninformationsdienst	IT	Tools	Vendor	Triumph-Adler
Institute	BSA	IT	TP	Vendor	Unidata
Institute	BSI	IT	Transaction	Vendor	Uniface
Institute	CE	IT	Transaktion	Vendor	Uniplex
Institute	CISR	IT	Transaktionen	Vendor	Unisource
Institute	Computergram	IT	Transformation	Vendor	Unisys

Dim	Term	Dim	Term	Dim	Term
Institute	CSC	IT	Treiber	Vendor	Univac
Institute	Dataquest	IT	Tutorials	Vendor	USL
Institute	Dean	IT	TV	Vendor	Varian
Institute	ESA	IT	uebertragen	Vendor	Vascom
Institute	Forrester	IT	uebertragung	Vendor	VEB
Institute	Gartner	IT	UIMS	Vendor	Vebacom
Institute	Gedas	IT	Umstieg	Vendor	Veritas
Institute	GMD	IT	UMTS	Vendor	Viag
Institute	HIS	IT	Universalrechner	Vendor	Vignette
Institute	IETF	IT	Unix-Rechner	Vendor	Vobis
Institute	IIR	IT	Update	Vendor	Vodafone
Institute	Integrata	IT	Updates	Vendor	Wang
Institute	IRD	IT	Upgrade	Vendor	Weblogic
Institute	KPMG	IT	Usern	Vendor	Worldcom
Institute	Lynch	IT	Variante	Vendor	Wyse
Institute	Merlin	IT	Varianten	Vendor	Xerox
Institute	Merrill	IT	verarbeiten	Vendor	Yahoo
Institute	OSF	IT	Verarbeitung		

Taxonomy Dim (Allianz):

Table 70: Taxonomy Dim (Allianz)

CC_Dim	Term	CC_Dim	Term	CC_Dim	Term
BusinessTerm	Abschreibungen	general	betraechtlich	general	selbst
BusinessTerm	AG	general	betraegt	general	setzte
BusinessTerm	Aktie	general	betreibt	general	sie
BusinessTerm	Aktien	general	betroffen	general	sind
BusinessTerm	Aktiengesellschaft	general	betrug	general	Situation
BusinessTerm	Aktienportefeuilles	general	betrugen	general	so
BusinessTerm	Aktionaeren	general	Bevoelkerung	general	soll
BusinessTerm	Altersversorgung	general	Beziehungen	general	sollte
BusinessTerm	angestellte	general	bezueglich	general	Sonstige
BusinessTerm	Angestellten	general	bietet	general	sonstigen
BusinessTerm	Anlage	general	Bild	general	Sorgen
BusinessTerm	Anteil	general	bis	general	sowie
BusinessTerm	Anteile	general	bisher	general	sowohl
BusinessTerm	Arbeit	general	bleibt	general	spaeter
BusinessTerm	arbeiten	general	blieb	general	Spezielle
BusinessTerm	Arbeitslosenquote	general	brachte	general	speziellen
BusinessTerm	Assets	general	brachten	general	spuerbar
BusinessTerm	aufgewendet	general	Buerger	general	spuerbaren
BusinessTerm	Aufwand	general	bzw	general	St
BusinessTerm	Aufwendungen	general	Da	general	Staaten
BusinessTerm	Aushilfen	general	Dabei	general	Stadt
BusinessTerm	Auslandsgeschaeft	general	Dachmontage	general	staendig
BusinessTerm	Auslandsgeschaefts	general	dadurch	general	staerker
BusinessTerm	Ausschuettung	general	dagegen	general	stammten
BusinessTerm	Auto-	general	daher	general	stand
BusinessTerm	Automatisierung	general	damit	general	stark
BusinessTerm	Baukonjunktur	general	Danach	general	starken
BusinessTerm	Behauptungsentzerrungsreserve	general	danken	general	starker
BusinessTerm	Beherrschungsvertraege	general	Darueber	general	steht
BusinessTerm	Beitraege	general	dass	general	steigende
BusinessTerm	Beitrag	general	Davon	general	steigern
BusinessTerm	Beitragsaufkommen	general	dazu	general	steigerte
BusinessTerm	Beitragseinnahmen	general	de	general	Steigerung
BusinessTerm	Beitragssteigerung	general	denen	general	Steigerungen
BusinessTerm	Beitragsvolumen	general	deren	general	steigt
BusinessTerm	Beitragswachstum	general	derselben	general	Stellen
BusinessTerm	Beitragszuwachs	general	Deshalb	general	stieg
BusinessTerm	Bericht	general	dessen	general	stiegen
BusinessTerm	Berichterstattung	general	deutlich	general	Struktur
BusinessTerm	Berichtsjahr	general	Dez	general	Summen
BusinessTerm	Berichtsjahres	general	Dies	general	Tabelle
BusinessTerm	Bestand	general	diese	general	taetigen
BusinessTerm	Bestandes	general	diesem	general	Taetigkeit
BusinessTerm	beteiligt	general	diesen	general	Technik
BusinessTerm	Beteiligung	general	dieser	general	Technische
BusinessTerm	Beteiligungen	general	Dieses	general	Technischen
BusinessTerm	Beteiligungsgesellschaft	general	diesmal	general	technischer
BusinessTerm	Bilanzen	general	differenzierter	general	Teil
BusinessTerm	Bilanzstichtag	general	direkte	general	Tendenz
BusinessTerm	Bilanzwert	general	direkten	general	Tendenzen
BusinessTerm	Branche	general	doch	general	This
BusinessTerm	Branchen	general	dort	general	Tier
BusinessTerm	brutto	general	Dr	general	totalen
BusinessTerm	Bruttobeitraege	general	drei	general	Trend
BusinessTerm	Bruttobeitragseinnahmen	general	Drittel	general	Trotz
BusinessTerm	Bruttopraemieneinkommen	general	drohende	general	trug
BusinessTerm	Co	general	durch	general	u.
BusinessTerm	Company	general	Durchschnitt	general	ueber
BusinessTerm	Controlling	general	durchschnittliche	general	ueberdurchschnittliche
BusinessTerm	Darlehen	general	durchschnittlichen	general	Uebereinstimmung
BusinessTerm	Depotforderungen	general	ebenfalls	general	Ueberpruefung
BusinessTerm	Direktionen	general	ebenso	general	ueberwiegend
BusinessTerm	Dividende	general	eigene	general	uebrige
BusinessTerm	DM-Rechnung	general	eigenen	general	Uebriges
BusinessTerm	Durchschnittsaufwand	general	Eigenschaft	general	um
BusinessTerm	Eigenkapitalinvestitionen	general	Eigenschaften	general	Umfang
BusinessTerm	Einbruchdiebstahl	general	Einbeziehung	general	umfasste

CC_Dim	Term	CC_Dim	Term	CC_Dim	Term
BusinessTerm	Einbruch-Diebstahl-Versicherung	general	einbezogene	general	Umfeld
BusinessTerm	eingezahlten	general	einbezogenen	general	Umstaende
BusinessTerm	Einkommen	general	Einblick	general	unbefriedigend
BusinessTerm	EK56	general	eine	general	unguenstig
BusinessTerm	Entnahme	general	einem	general	unguenstigen
BusinessTerm	Erfolg	general	einen	general	unguenstigeren
BusinessTerm	erfolgreichen	general	einer	general	uns
BusinessTerm	Ergebnis	general	eines	general	Unser
BusinessTerm	Ergebnisse	general	Einfluesse	general	unsere
BusinessTerm	Ergebnisverbesserung	general	Einfluss	general	unserem
BusinessTerm	Ertraege	general	Einfuehrung	general	unseren
BusinessTerm	Ertrag	general	eingerichtet	general	unserer
BusinessTerm	Ertragslage	general	eingetreten	general	unseres
BusinessTerm	Ertragsverbesserung	general	einige	general	unter
BusinessTerm	erwarb	general	einigen	general	unterschiedlich
BusinessTerm	Erwerbe	general	einiger	general	unterstuetzt
BusinessTerm	erworbenen	general	einmal	general	Ursachen
BusinessTerm	erzielt	general	einschliessend	general	Veraenderung
BusinessTerm	erzielten	general	einschliesslich	general	Veraenderungen
BusinessTerm	Festverzinsliche	general	einzelne	general	verbessert
BusinessTerm	FINANZDIENSTLEISTUNGEN	general	einzelnen	general	verbesserten
BusinessTerm	Finanzjahr	general	einzige	general	Verbesserung
BusinessTerm	Fireman	general	Ende	general	Verbindung
BusinessTerm	Forderung	general	entfielen	general	verbunden
BusinessTerm	Garantie-	general	enthalten	general	Verbundene
BusinessTerm	garantieren	general	Entsprechend	general	Vereinigten
BusinessTerm	Gesamtergebnis	general	entspricht	general	Vereinte
BusinessTerm	Gesamtsumme	general	Entwicklung	general	Verfahren
BusinessTerm	gesamtwirtschaftlichen	general	Entwicklungen	general	Verfassung
BusinessTerm	Geschaeft	general	er	general	vergangenen
BusinessTerm	Geschaeftes	general	ergab	general	Vergleich
BusinessTerm	Geschaefts	general	ergibt	general	Verhaeltnis
BusinessTerm	Geschaeftsausweitung	general	erhalten	general	verhaeltnismaessig
BusinessTerm	Geschaeftsjahr	general	erheblich	general	Verhaeltnisse
BusinessTerm	Geschaeftsjahres	general	Erhebliche	general	Verlauf
BusinessTerm	Geschaeftspolitik	general	erheblichen	general	verlief
BusinessTerm	Geschaeftsverlauf	general	erhielt	general	vermehrten
BusinessTerm	Geschaeftsvolumen	general	erhielten	general	Verordnung
BusinessTerm	Gesellschaft	general	erhoehen	general	verpflichten
BusinessTerm	Gesellschaften	general	erhoeht	general	Verpflichtung
BusinessTerm	gesetzliche	general	erhoehte	general	verschieden
BusinessTerm	gesetzlichen	general	erhoehten	general	verschiedenen
BusinessTerm	gewerblichen	general	Erhoehung	general	verstaerken
BusinessTerm	Gewinn	general	erkennen	general	verstaerkt
BusinessTerm	Gewinn-	general	erklaert	general	verstaerkten
BusinessTerm	Gewinnanteils	general	ermoeglicht	general	Verteilung
BusinessTerm	Gewinnbeteiligung	general	erneut	general	verwendet
BusinessTerm	GmbH	general	Ernst	general	verzeichnen
BusinessTerm	Grossschaeden	general	erreicht	general	viel
BusinessTerm	Group	general	erreichte	general	viele
BusinessTerm	Gruendung	general	erreichten	general	Viertel
BusinessTerm	Gruppe	general	erste	general	Vj
BusinessTerm	Gruppengesellschaften	general	ersten	general	voll
BusinessTerm	gutgeschrieben	general	erster	general	vollstaendig
BusinessTerm	Haftpflicht	general	erstmalige	general	Volumen
BusinessTerm	Haftpflichtversicherer	general	erstmals	general	vom
BusinessTerm	Haftpflichtversicherung	general	erwarten	general	vor
BusinessTerm	Haftung	general	erwartet	general	vorhandene
BusinessTerm	Handel	general	Es	general	Vorjahr
BusinessTerm	Holding	general	et	general	Vorjahren
BusinessTerm	Hypotheken	general	etwa	general	Vorjahres
BusinessTerm	IAS	general	etwas	general	Vorwiegend
BusinessTerm	Inc	general	exkl	general	Waehrend
BusinessTerm	Industriegeschaeft	general	exklusiv	general	war
BusinessTerm	Industriekunden	general	f-	general	waren
BusinessTerm	industrielle	general	Faktor	general	Was
BusinessTerm	industriellen	general	fast	general	Weg
BusinessTerm	industrieller	general	Februar	general	wegen
BusinessTerm	Inflationsrate	general	Feld	general	weil
BusinessTerm	Inlandsgeschaeft	general	Fernen	general	Weise
BusinessTerm	Insurance	general	ferner	general	weit
BusinessTerm	Investition	general	feste	general	weiter
BusinessTerm	Investitionen	general	fiel	general	weitere
BusinessTerm	Investitionseinkommen	general	Folge	general	weiteren
BusinessTerm	Jahresabschluss	general	Folgen	general	weiterhin
BusinessTerm	Jahresueberschuss	general	folgende	general	welchem
BusinessTerm	Joint	general	Form	general	welches
BusinessTerm	Kapitalanlageergebnis	general	fort	general	wenig
BusinessTerm	Kapitalanlagen	general	fortgesetzt	general	weniger
BusinessTerm	kapitalbildenden	general	freien	general	Wenn
BusinessTerm	Kapitalerhoehungen	general	frueh	general	wer
BusinessTerm	Kapitalertraege	general	frueheren	general	werdenden
BusinessTerm	Konsolidierte	general	fuehlt	general	wesentlich
BusinessTerm	konsolidierten	general	fuehren	general	wesentliche
BusinessTerm	Konsolidierung	general	Fuehrender	general	Wesentlichen
BusinessTerm	Konsortium	general	fuehrte	general	wichtige
BusinessTerm	Konzern	general	fuehrten	general	wichtigen
BusinessTerm	Konzernabschluss	general	Fuehrung	general	wie
BusinessTerm	Konzerns	general	fuenf	general	wieder
BusinessTerm	Konzernunternehmen	general	ganze	general	wiederum
BusinessTerm	Kosten	general	ganzen	general	wir
BusinessTerm	Kostenquote	general	geben	general	wird
BusinessTerm	Kostensteigerungen	general	gefuehrt	general	wirklicher

CC_Dim	Term	CC_Dim	Term	CC_Dim	Term
BusinessTerm	Kostenverhaeltnis	general	gegeben	general	wirkt
BusinessTerm	Kreditrisiken	general	gegen	general	wirkte
BusinessTerm	Kunde	general	gegenueber	general	Wirkung
BusinessTerm	Kunden	general	gehalten	general	wissen
BusinessTerm	Kurs	general	gehaltenen	general	wo
BusinessTerm	Lagebericht	general	gehandelt	general	worden
BusinessTerm	Lebenshaltung	general	gehoert	general	wuchs
BusinessTerm	Lebenshaltungskosten	general	geht	general	wuerde
BusinessTerm	Liefern	general	geltenden	general	wurde
BusinessTerm	Lohn-	general	gemacht	general	wurden
BusinessTerm	Ltd	general	generierte	general	Zahl
BusinessTerm	Maerkte	general	genommen	general	Zahlen
BusinessTerm	Management	general	gerade	general	zeichnete
BusinessTerm	MANAGEMENT/FINANZDIENSTLEISTUNGEN	general	geraten	general	zeichneten
BusinessTerm	Markt	general	geringere	general	zeigt
BusinessTerm	Markt-	general	geringerer	general	zeigte
BusinessTerm	Marktfuehrer	general	gesamten	general	zeigten
BusinessTerm	Marktposition	general	Geschichten	general	Zeit
BusinessTerm	Markts	general	geschrieben	general	Zentrum
BusinessTerm	Maschinen	general	geschriebene	general	ziemlich
BusinessTerm	mbH	general	gesteigert	general	Ziff
BusinessTerm	Mitarbeiter	general	gesteigerter	general	zudem
BusinessTerm	Mitarbeiterinnen	general	gestiegen	general	zuerst
BusinessTerm	Mitarbeitern	general	gestiegene	general	zufrieden
BusinessTerm	Mitglieder	general	gestiegenen	general	zufriedenstellend
BusinessTerm	Muttergesellschaft	general	Gestuetzt	general	zufriedenstellender
BusinessTerm	Nationalwirtschaften	general	gewesen	general	Zufuehrung
BusinessTerm	netto	general	gezeigt	general	Zugenommen
BusinessTerm	Nettoeinkommen	general	gibt	general	zugute
BusinessTerm	Nettopraemien	general	gilt	general	Zukunft
BusinessTerm	Neugeschaeft	general	ging	general	zuletzt
BusinessTerm	nichtversicherungstechnischen	general	gingen	general	zum
BusinessTerm	Nutzen	general	gleiche	general	Zunahme
BusinessTerm	operative	general	gleichen	general	zunehmend
BusinessTerm	organisatorischen	general	globale	general	zunehmende
BusinessTerm	Personal-	general	globalen	general	zunehmenden
BusinessTerm	plc	general	Grenzen	general	zur
BusinessTerm	Portefeuille	general	groesser	general	zurueck
BusinessTerm	Portefeuilles	general	groesseren	general	zurueckgegangen
BusinessTerm	Position	general	groesste	general	zurueckzufuehren
BusinessTerm	Praemien	general	grosse	general	ZurVerfuegung
BusinessTerm	Praemienaufkommen	general	Grossen	general	Zusaetzlich
BusinessTerm	Praemienaufkommens	general	grosser	general	zusammen
BusinessTerm	Praemieneinkommen	general	Grund	general	Zusammenarbeit
BusinessTerm	Praemieneinnahme	general	guenstig	general	Zusammenfassung
BusinessTerm	Praemieneinnahmen	general	Guenstige	general	Zusammenhang
BusinessTerm	Praemiensaetze	general	gut	general	zuschreibbar
BusinessTerm	Praemienvolumen	general	gute	general	zustaendigen
BusinessTerm	Praemienwachstum	general	gutem	general	Zuwachs
BusinessTerm	Preis	general	guten	general	zuzunehmen
BusinessTerm	Produkte	general	haben	general	zwar
BusinessTerm	Projekt	general	halten	general	Zweck
BusinessTerm	Rationalisierung	general	Hand	general	zwei
BusinessTerm	Rechnung	general	harten	general	Zweiten
BusinessTerm	Regionalstruktur	general	hat	general	zwischen
BusinessTerm	Regulierung	general	hatte	Geography	Afrika
BusinessTerm	reorganisieren	general	hatten	Geography	America
BusinessTerm	Reserven	general	Hauptgrund	Geography	American
BusinessTerm	RM	general	hauptsaechlich	Geography	Amerika
BusinessTerm	Ruecklage	general	Herr	Geography	amerikanischen
BusinessTerm	Ruecklagen	general	heute	Geography	Amsterdam
BusinessTerm	Rueckstellung	general	hielt	Geography	Asien
BusinessTerm	Rueckstellungen	general	hier	Geography	auslaendische
BusinessTerm	Ruhestandsvorsorge	general	Hierfuer	Geography	auslaendischen
BusinessTerm	s Fund	general	Hierin	Geography	Ausland
BusinessTerm	S.A	general	hiervon	Geography	Australien
BusinessTerm	Sanierung	general	Hilfe	Geography	Bayerische
BusinessTerm	Sanierungen	general	hinauf	Geography	Bayerischen
BusinessTerm	Sanierungsmassnahmen	general	Hinblick	Geography	Bayern
BusinessTerm	Satzung	general	hingewiesen	Geography	Brasilien
BusinessTerm	Schwankungsrueckstellung	general	Hoehe	Geography	Bremen
BusinessTerm	Schwankungsrueckstellungen	general	hoeher	Geography	britischen
BusinessTerm	Service	general	hoehere	Geography	Bundeslaendern
BusinessTerm	Sicherheiten	general	hoeheren	Geography	Bundesrepublik
BusinessTerm	Stammkapital	general	hohe	Geography	Chile
BusinessTerm	Steigerungsraten	general	hohen	Geography	Deutsche
BusinessTerm	Steuern	general	ich	Geography	deutschen
BusinessTerm	strategischen	general	ihm	Geography	Deutsches
BusinessTerm	Tarif	general	ihn	Geography	Deutschland
BusinessTerm	Tarife	general	ihr	Geography	Europa
BusinessTerm	Teilabschreibungen	general	ihre	Geography	europaeische
BusinessTerm	Tochtergesellschaft	general	ihren	Geography	Europaeischen
BusinessTerm	Tochtergesellschaften	general	ihrer	Geography	Europas
BusinessTerm	Transport	general	II	Geography	Frankfurt
BusinessTerm	Uebernahme	general	immer	Geography	Frankfurter
BusinessTerm	Uebernahmeangebot	general	indirekt	Geography	Frankfurts
BusinessTerm	uebernommen	general	indirekten	Geography	Frankreich
BusinessTerm	uebernommene	general	Infolge	Geography	Frankreichs
BusinessTerm	uebernommenen	general	insbesondere	Geography	franzoesische
BusinessTerm	Ueberschuss	general	Insgesamt	Geography	franzoesischen
BusinessTerm	Umsatz	general	Interesse	Geography	Grossbritannien
BusinessTerm	Unternehmen	general	Interessen	Geography	Hamburg
BusinessTerm	Unternehmens	general	jaehrige	Geography	Hungaeria

CC_Dim	Term	CC_Dim	Term	CC_Dim	Term
BusinessTerm	US-Dollar	general	Jahr	Geography	Indonesia
BusinessTerm	US-Dollars	general	Jahre	Geography	Indonesien
BusinessTerm	Verdienst	general	Jahren	Geography	indonesische
BusinessTerm	Verkaeufe	general	Jahres	Geography	inlaendischen
BusinessTerm	Verkauf	general	Jahresdurchschnitt	Geography	Inland
BusinessTerm	verkauft	general	Januar	Geography	Inlands-
BusinessTerm	Verlust	general	je	Geography	International
BusinessTerm	Verluste	general	jede	Geography	internationale
BusinessTerm	Verlustrechnung	general	jedes	Geography	internationalen
BusinessTerm	Verlustsituation	general	Jedoch	Geography	Italien
BusinessTerm	Verlustverhaeltnis	general	jenen	Geography	italienischen
BusinessTerm	Vermoegensanlagen	general	jetzt	Geography	Japan
BusinessTerm	Vertraege	general	jeweiligen	Geography	japanischen
BusinessTerm	Vertraegen	general	jeweils	Geography	Johannesburg
BusinessTerm	Vertrag	general	Kalender	Geography	Kanada
BusinessTerm	Verwaltungsgesellschaft	general	kam	Geography	Karlsruher
BusinessTerm	Volkswirtschaft	general	kann	Geography	lokalen
BusinessTerm	Vorsorge	general	keine	Geography	London
BusinessTerm	Vorstand	general	keineswegs	Geography	Mexico
BusinessTerm	VORSTANDS	general	Klasse	Geography	Mexiko
BusinessTerm	Vortrag	general	Klassen	Geography	Mexikos
BusinessTerm	Wachstum	general	knapp	Geography	Mosel
BusinessTerm	Waehrungen	general	Know-how	Geography	Muenchen
BusinessTerm	Wert	general	koennen	Geography	Muenchener
BusinessTerm	Wertminderung	general	kommt	Geography	Oesterreich
BusinessTerm	Wertpapiere	general	Kompetenz	Geography	oesterreichischen
BusinessTerm	Wettbewerb	general	konnte	Geography	Polen
BusinessTerm	Wettbewerbs	general	konnten	Geography	Portugal
BusinessTerm	Wirtschaft	general	kontrolliert	Geography	Singapur
BusinessTerm	wirtschaftlichen	general	Kraeftig	Geography	Spanien
BusinessTerm	Wirtschaftsentwicklung	general	Kraft	Geography	Stuttgart
BusinessTerm	Wirtschaftsgemeinschaft	general	Kraftfahrt	Geography	Suedafrika
BusinessTerm	Zahlungsverpflichtungen	general	Kraftfahrzeugtechnik	Geography	suedafrikanischen
BusinessTerm	Zeichnungsgewinn	general	Krankenkassen	Geography	Suedkorea
BusinessTerm	Zeichnungspolitik	general	kuenftig	Geography	Sydney
BusinessTerm	Zinssaetze	general	Laendern	Geography	Thailand
BusinessTerm	Zuwachsquote	general	laesst	Geography	Togo
BusinessTerm	Zweige	general	lag	Geography	Ungarn
BusinessTerm	Zweigen	general	Lage	Geography	US-
BusinessTerm	Zweigniederlassungen	general	Landes	Geography	USA
BusinessTerm	Zweigstelle	general	Landwirte	Geography	Welt
BusinessTerm	Zweigstellen	general	landwirtschaftlichen	Geography	weltweit
Company	Adriatico	general	langfristige	Geography	Wiener
Company	AGF	general	langjaehrig	InsuranceTerm	Erdbeben
Company	Allianz	general	lassen	InsuranceTerm	Fahrzeuge
Company	Allianz-Gesellschaften	general	laufende	InsuranceTerm	Feuer
Company	Allianz-Gruppe	general	laufenden	InsuranceTerm	Feuer-
Company	Allianzs	general	Leben	InsuranceTerm	Feuerversicherung
Company	Allianz-Sachgruppe	general	Lebens-	InsuranceTerm	Feuerwehrmann
Company	Assecuranz-Compagnie	general	lediglich	InsuranceTerm	Feuerweiterversicherung
Company	Assurances	general	leicht	InsuranceTerm	Frostperioden
Company	AZT	general	leichter	InsuranceTerm	Glas
Company	Cornhill	general	Leistung	InsuranceTerm	Grundbesitz
Company	Datastream	general	Leistungen	InsuranceTerm	Grundstuecke
Company	Eagle	general	Leitung	InsuranceTerm	Grundvermoegen
Company	ELVIA	general	letzten	InsuranceTerm	Hagel
Company	Federales	general	levels	InsuranceTerm	Hagelsturm
Company	Globus	general	liegt	InsuranceTerm	Hausrat
Company	Hamburg-Mannheimer	general	Life	InsuranceTerm	Hausratversicherung
Company	Hermes	general	Linie	InsuranceTerm	Immobilien
Company	Lebensversicherungs-AG	general	lllllllll	InsuranceTerm	Industriefeuerversicherung
Company	Mercur	general	machen	InsuranceTerm	Industrieversicherung
Company	RAS	general	machte	InsuranceTerm	Insassen-Unfallversicherung
Company	Rechtsschutzversicherungs-AG	general	machten	InsuranceTerm	Kfz-Versicherung
Company	Rhin	general	Mal	InsuranceTerm	Kraftfahr-Haftpflichtversicherung
Company	Rueckversicherungs-Gesellschaft	general	Man	InsuranceTerm	Kraftfahrtversicherung
Company	Sicurtae	general	Mann	InsuranceTerm	Kraftfahrversicherung
Company	Star	general	Mass	InsuranceTerm	Kraftverkehrs-Strafrechtsschutz
Company	Treuegemeinschaftslebensversicherungsgesellschaft	general	Masse	InsuranceTerm	Krankenrueckversicherung
Company	Union	general	massgeblich	InsuranceTerm	Krankenversicherung
Company	Veritas	general	Massnahmen	InsuranceTerm	Krankenversicherungs
Company	Ver-sicherungs-AG	general	mehr	InsuranceTerm	Kreditversicherung
Currency	DM	general	mehren	InsuranceTerm	Kreditweiterversicherung
Currency	Euro	general	meiste	InsuranceTerm	Lebensversicherung
general	Abdeckung	general	meisten	InsuranceTerm	Lebensversicherungen
general	Aber	general	mittlere	InsuranceTerm	Leitungswasser
general	abgeschlossen	general	mn	InsuranceTerm	Leitungswasserschaeden
general	abgeschlossene	general	moeglich	InsuranceTerm	Luftfahrtversicherung
general	abgeschlossenen	general	Monat	InsuranceTerm	Maschinenversicherung
general	abgeschlossenes	general	Motor	InsuranceTerm	Naturkatastrophen
general	Abschluss	general	muessen	InsuranceTerm	Nicht-Leben
general	Abschlusses	general	muss	InsuranceTerm	Personenschaeden
general	Abwicklung	general	musste	InsuranceTerm	Rechtsschutzversicherung
general	abzugeben	general	nach	InsuranceTerm	Rechtsschutzversicherungs
general	aehnlichen	general	Nahezu	InsuranceTerm	Reiseversicherung
general	Aenderung	general	nahm	InsuranceTerm	Reparaturkosten
general	all	general	nahmen	InsuranceTerm	Risiken
general	alle	general	Name	InsuranceTerm	Risikenzahl
general	allein	general	neben	InsuranceTerm	Risiko
general	allem	general	Negative	InsuranceTerm	Risikogruppen
general	allen	general	Nennenswerte	InsuranceTerm	Risikomanagement

CC_Dim	Term	CC_Dim	Term	CC_Dim	Term
general	aller	general	neu	InsuranceTerm	Risikosituation
general	allerdings	general	neue	InsuranceTerm	Rueck
general	Allgemeine	general	neuen	InsuranceTerm	Rueckdeckung
general	allgemeinen	general	neuer	InsuranceTerm	Rueckgewaehrdauer
general	alten	general	nicht	InsuranceTerm	Rueckversicherung
general	am	general	niedrigeren	InsuranceTerm	Sachgruppe
general	an	general	nmen	InsuranceTerm	Sachversicherung
general	andere	general	noch	InsuranceTerm	Sachversicherungs
general	anderen	general	nochmalige	InsuranceTerm	Sachversicherungsgruppe
general	Andererseits	general	nochmals	InsuranceTerm	Schaden
general	Aneignung	general	nom	InsuranceTerm	Schaden-
general	Anfang	general	normale	InsuranceTerm	Schadenaufwand
general	Angelegenheit	general	notwendig	InsuranceTerm	Schadenaufwandes
general	angenommen	general	nur	InsuranceTerm	Schadenaufwands
general	Anhebung	general	o>	InsuranceTerm	Schadenaufwendungen
general	Anlauf	general	Obwohl	InsuranceTerm	Schadenbelastung
general	Anpassung	general	oder	InsuranceTerm	Schadendurchschnitt
general	Anreize	general	ohne	InsuranceTerm	Schadenentwicklung
general	Anstieg	general	Parlament	InsuranceTerm	Schadenfall
general	Anstiegs	general	physischen	InsuranceTerm	Schadenhaeufigkeit
general	anwachsen	general	ploetzlich	InsuranceTerm	Schadenleistungen
general	Anzahl	general	Politiken	InsuranceTerm	Schadenquote
general	Aufgabe	general	Populaer	InsuranceTerm	Schadenursachen
general	aufgefangen	general	positiv	InsuranceTerm	Schadenverlauf
general	Aufgrund	general	positive	InsuranceTerm	Schaeden
general	aus	general	positiven	InsuranceTerm	Selbstbeteiligung
general	Ausbau	general	positives	InsuranceTerm	Sparte
general	ausgeglichen	general	praktisch	InsuranceTerm	Sparten
general	ausgewiesen	general	President	InsuranceTerm	Stadions
general	ausgewirkt	general	private	InsuranceTerm	Sturm
general	Ausgleich	general	privaten	InsuranceTerm	Teilschaden
general	Ausmass	general	pro	InsuranceTerm	Tierversicherung
general	ausschlaggebend	general	Professor	InsuranceTerm	Transportversicherung
general	Ausserdem	general	Prozent	InsuranceTerm	Unfaellen
general	Ausweitung	general	Prozentsatz	InsuranceTerm	Unfall
general	auswirken	general	Quelle	InsuranceTerm	Unfallversicherung
general	Auswirkungen	general	Rahmen	InsuranceTerm	Unfallversicherungsgeschaeft
general	auszugleichen	general	Rechtsprechung	InsuranceTerm	Verbrechen
general	bald	general	reduzieren	InsuranceTerm	Versicherer
general	Baustelle	general	Region	InsuranceTerm	versicherten
general	Bedeutung	general	Regionen	InsuranceTerm	Versicherung
general	bedingt	general	reichte	InsuranceTerm	Versicherungen
general	bedingte	general	Republik	InsuranceTerm	Versicherungs
general	beeinflusst	general	Rest	InsuranceTerm	Versicherungs-AG
general	beeinflussten	general	Rolle	InsuranceTerm	Versicherungsbank
general	beeintraechtigt	general	Rueckgang	InsuranceTerm	Versicherungsgeschaeft
general	befindet	general	rund	InsuranceTerm	Versicherungsgeschaefts
general	befriedigend	general	rung	InsuranceTerm	Versicherungsgesellschaft
general	begann	general	s core	InsuranceTerm	Versicherungsindustrie
general	Behauptungen	general	S.	InsuranceTerm	Versicherungsmarkt
general	bei	general	Saemtliche	InsuranceTerm	Versicherungsnehmer
general	beide	general	sah	InsuranceTerm	Versicherungsnehmern
general	beiden	general	scharf	InsuranceTerm	Versicherungsschutz
general	beigetragen	general	schliesst	InsuranceTerm	Versicherungssumme
general	beim	general	schloessen	InsuranceTerm	Versicherungssummen
general	Beispiel	general	schloss	InsuranceTerm	Versicherungstechnische
general	Beispiele	general	Schnitt	InsuranceTerm	versicherungstechnischen
general	Belastung	general	schon	InsuranceTerm	Versicherungsunternehmen
general	Belebung	general	schreiben	InsuranceTerm	Versicherungswirtschaft
general	Bemuehungen	general	Schutz	InsuranceTerm	Versicherungszweige
general	Bereich	general	Schwankungen	InsuranceTerm	Versicherungszweigen
general	bereitet	general	Schwerpunkt	InsuranceTerm	Vertreter
general	bereits	general	sechs	InsuranceTerm	Weiterversicherung
general	beschaeftigten	general	Seeplatz	Name	Bartholomew
general	besondere	general	sehr	Name	Benediktbeurer
general	Besonderen	general	sein	Name	Euler
general	besonders	general	seine	Name	Jones
general	besser	general	seinem	Name	Kornai
general	bestehen	general	seinen	Name	Thomas
general	besteht	general	seit		

Semantical Analysis CW_{5k}, leading aggregated concepts within corpus segments and drill-down to term level:

Table 71: CW_{5k}, leading aggregated concepts within corpus segments and drill-down to term level

	CW_{5k}		
Date	1975	1988	2003
Cc_CountThresU_Dim	Currency	Currency	Currency
	Vendor	Vendor	Economy
	Profession	Economy	Vendor
	IT	IT	Geography
	Geography	Customer	IT
	Economy	Business	Business
	OS	Geography	Performance
	Customer	Profession	Profession
	Business	Performance	Science
	Performance	Science	Customer
	Science	Event	Name
Cv_CountThresU_Dim	Currency	OS	Currency
	Vendor	Norm	OS
	IT	Institute	ProgLanguage
	ITProduct	Vendor	Vendor
	OS	ITProduct	Profession
	ProgLanguage	ProgLanguage	Institute
	Performance	IT	IT
	Institute	Currency	ITProduct
	Profession	Event	Economy
	Name	Performance	Performance
	Norm	Science	Customer
	Business	SocialFramework	Name
	Economy	Customer	Business
	SocialFramework	Economy	Geography
	Customer	Geography	Norm
	Geography	Profession	SocialFramework
	Event	Business	Event
	Science	Name	Science
Chipcom.TermFirstOcc	1987		
Chipcom.TermLastOcc	1998		
Cc_CountThresU_Vendor	IBM	IBM	IBM
	Siemens	DEC	HP
	Nixdorf	Siemens	Intel
	Bull	IBMs	Siemens
	Philips	Digital	Hewlett-Packard
	Xerox	HP	
	Digital	Hewlett-Packard	
	NCR	Bull	
		Nixdorf	
		NCR	
Cv_CountThresU_Vendor	Honeywell	Sun	Microsoft
	Univac	Apple	SAP
	Unidata	Microsoft	Sun
	Kienzle	Bundespost	Oracle
	Bundespost	Novell	SCO
	CDC	Unisys	Abb
	Burroughs	Oracle	Microsofts
	Singer	Wang	Telekom
	Sperry	Apollo	Suse
	Interdata	Amdahl	EDS
	Amdahl	Systec	Cisco
	Datev	Ashton-Tate	Novell
	ADV/ORGA	3Com	Google
	Memorex	Compaq	Siebel
	Compagnie	Toshiba	Bea
	Hewlett	Sony	Intels
	MDS	Kombinat	Symantec
	Taylorix	Atlantic	Weblogic
	MBB	Datev	FSC
	Facit	SCO	T-Mobile
	Packard	DECs	EMC
	Varian	NEC	Ericsson
	Inforex	Nokia	Palm
	SEL	EDS	Vignette
	Triumph-Adler	Vascom	Vodafone
	Calcomp	Harris	Apple
	Datasaab	VEB	Navision
	AEG-Telefunken	Informix	Sybase
	Olympia	AEG	AMD

	CW_{5k}		
Date	1975	1988	2003
	Wang	ADR	AOL
	Anker	IDV	Ixos
	CTM	Robotron	Suns
	AEG	Tandem	Borland
	GMO	CDC	SGI
	IBM-	Kodak	Sony
	ITT	Norsk	Unisys
	MS	Stratus	Infineon
	Centronics	Honeywell	Compaq
	Adler	Kienzle	Fujitsu-Siemens
	Harris	Novells	IT-Anbieter
	SAP	Sybase	T-Online
	AMD	Cullinet	Lexmark
	SPC	Epson	Mobilcom
	Bosch	SEL	Nokia
	Kodak	Alcatel	SAS
	Bayer	Commodore	Documentum
	DBP	Convex	Netscape
	Robotron	Ericsson	Oracles
	Mergard	IBM-	HPs
	Sprint	Microsofts	3Com
	Tektronix	Mitsubishi	BMW
		AMD	Acer
		Olympia	Toshiba
		Bayer	Amazon
		Acer	Adobe
		Apples	Corel
		Datapoint	Gates-Company
		DBP	MCI
		MS	Cognos
		STC	SAPs
		Abb	Veritas
		Ameritech	Intergraph
		Atari	Intershop
		BMW	Lucent
		Bosch	MSN
		Comparex	NEC
		Computerland	Ariba
		CTM	Samsung
		Cunningham	Debis
		Parsytec	Altavista
		Triumph-Adler	Baan
		SAS	Bayer
		Tandon	Compunet
		Bertelsmann	Hyperion
		HPs	Seagate
		Kontron	Freenet
		MCI	Heiler
		Prologue	Alcatel
		Retix	Sharp
		Sharp	Lycos
		Suns	Microstrategy
		Digitals	Nortel
		Intels	Novells
		MBB	Bosch
		Memorex	Apollo
		Oracles	Apples
		Packard	Intuit
		Sequent	KPN
		SNI	McAfee
		Sperry	Tibco
		Taylorix	Atari
		Adobe	Atlantic
		Cabletron	Bertelsmann
		Inmos	IBM-
		Intergraph	Mitsubishi
		ITT	Pixelpark
		SAP	Seebeyond
		Telekom	Wang
		Unidata	Wyse
		Baan	
		BT	
		Burroughs	
		Centronics	
		Compunet	
		Hewlett	
		IBM-Tochter	
		Matsushita	
		Symantec	
		Tektronix	

	CW_{5k}		
Date	1975	1988	2003
		Wyse	

Semantical Analysis CW_{1K}, leading aggregated concepts within corpus segments and drill-down to term level:

Table 72: CW_{1k}, leading aggregated concepts within corpus segments and drill-down to term level

	CW_{1k}		
Date	1975	1988	2003
Cc_CountThresU_Dim	Vendor	Vendor	Currency
	Currency	Currency	Economy
	IT	IT	Geography
	Business	Economy	Vendor
	Economy	Business	IT
	Profession	Geography	Business
	Geography	Customer	
	Customer		
Cv_CountThresU_Dim	OS	OS	Currency
	Vendor	Vendor	OS
	Customer	Event	Vendor
	Institute	IT	ProgLanguage
	IT	ITProduct	Institute
	ITProduct	Geography	Profession
	Currency	Norm	Customer
	Economy	Performance	Economy
	Geography	Science	IT
	Name	Business	ITProduct
	Profession	Currency	Norm
	Science	Customer	Business
	Business	Economy	Name
	Event	Institute	Performance
	Norm	Profession	Geography
	ProgLanguage	SocialFramework	Science
	SocialFramework	Name	SocialFramework
	Performance	ProgLanguage	Event
Chipcom.TermFirstOcc	1989		
Chipcom.TermLastOcc	1996		
Cc_CountThresU_Vendor	IBM	IBM	IBM
	Siemens	Siemens	HP
Cv_CountThresU_Vendor	Unidata	DEC	SAP
	Nixdorf	Sun	Microsoft
	Honeywell	Apple	Sun
	Univac	Intel	Dell
	BASF	Microsoft	Oracle
	Olivetti	Oracle	SCO
	Xerox	Bull	Suse
	Burroughs	Nixdorf	Lexmark
	Kienzle	Fujitsu	Siebel
	Bull	Toshiba	Microsofts
	CDC	Motorola	Gemini
	Singer	Olivetti	Vodafone
	Memorex	Unisys	Apple
	mbp	DECs	Bull
	Sperry	EDS	EMC
	Mannesmann	Novell	Mobilcom
	Taylorix	Apollo	Cisco
	Wang	Atlantic	EDS
	Amdahl	Tandem	Novell
	Compagnie	Wang	Sony
	Bundespost	Ashton-Tate	Suns
	Centronics	Bell	T-Mobile
	Olympia	Ericsson	Compaq
	MS	Harris	Compunet
	Triumph-Adler	SEL	FSC
	Diebold	Xerox	Nokia
	MBB	Bundespost	Unisys
	AEG-Telefunken	Hitachi	Freenet
	Anker	Semiconductor	Intel
	DEC	Microsofts	Intershop
	Facit	NEC	SGI
	Interdata	SCO	Telekom
	Mergard	Commodore	Vignette
	Sprint	Compaq	Mannesmann
	Bayer	Epson	Motorola
	DBP	Gemini	NEC
	Fujitsu	Kienzle	Adobe

	CW_{1k}		
Date	1975	1988	2003
	GMO	SNI	Fujitsu-Siemens
	Hewlett	Stratus	Ixos
	IBM-	Suns	Weblogic
	Inforex	Systec	Xerox
	Matsushita	3Com	Debis
	Packard	Amdahl	Navision
	Robotron	Atari	Novells
		Cray	SAPs
		CTM	SAS
		Dell	Symantec
		Mannesmann	Toshiba
		SAS	3Com
		Softlab	BMW
		Sony	Fujitsu
		Unidata	Heiler
		Alcatel	Hitachi
		Apples	Intels
		Bertelsmann	KPN
		BMW	MSN
		Computerland	Oracles
		DECnet	Samsung
		GE	Softlab
		Honeywell	Wyse
		IDV	Amazon
		Informix	AMD
		Intels	BASF
		Nokia	Bayer
		Oracles	Bosch
		Sequent	Cray
		Sybase	Documentum
		Taylorix	Ericsson
			Gates-Company
			GE
			Hyperion
			IT-Anbieter
			Lucent
			Netscape
			Nortel
			Palm
			T-Online

Semantical Analysis CW_{5Kb}, leading aggregated concepts within corpus segments and drill-down to term level:

Table 73: CW_{5kb}, leading aggregated concepts within corpus segments and drill-down to term level

	CW_{5kb}		
Date	1975	1988	2003
Cc_CountThresU_Dim	IT	IT	IT
	Science	Vendor	Currency
	Currency	Science	Profession
	Vendor	SocialFramework	Vendor
	Business	Currency	Science
	Profession	Profession	Economy
	Economy	Economy	Event
	SocialFramework	Business	Business
	Geography	Performance	Geography
	Customer	Name	
		Geography	
		Customer	
Cv_CountThresU_Dim	Event	Vendor	Currency
	OS	ProgLanguage	Norm
	Name	Profession	OS
	Vendor	Name	Profession
	Institute	OS	Customer
	Business	ITProduct	ITProduct
	Geography	IT	Vendor
	Economy	Institute	ProgLanguage
	IT	Business	IT
	Currency	Norm	Economy
	Profession	Geography	Institute
	ITProduct	Performance	Geography
	Customer	Economy	Business
	Norm	Science	Name
	Performance	Currency	SocialFramework
	Science	SocialFramework	Performance
	SocialFramework	Customer	Event
	ProgLanguage	Event	
Chipcom.TermFirstOcc	1987		
Chipcom.TermLastOcc	1998		
Cc_CountThresU_Vendor	IBM	IBM	IBM
	Microsoft	Microsoft	Siemens
	Siemens	Fujitsu	Microsoft
	HP	DEC	HP
	Apple	Apple	Sharp
	Nixdorf	Digital	SAP
		SAP	Intel
		Siemens	
		HP	
Cv_CountThresU_Vendor	Telekom	Telekom	Novell
	3Com	Bertelsmann	Oracle
	Vodafone	Novell	SCO
	Infineon	SAS	Vodafone
	Oracle	Hyperion	Borland
	Ariba	Infineon	Fujitsu-Siemens
	Lycos	Toshiba	Abb
	T-Online	Mobilcom	Microsofts
	AMD	Siebel	Telekom
	Matsushita	Oracle	EDS
	Nokia	Pixelpark	Suse
	Microstrategy	Nokia	Siebel
	Worldcom	SAPs	Cisco
	Honeywell	AOL	Bea
	Univac	Freenet	Google
	Unidata	Bundespost	Intels
	Kienzle	Amdahl	Symantec
	Sperry	Unisys	Ericsson
	Bundespost	Wang	Weblogic
	Burroughs	Apollo	EMC
	Singer	Ashton-Tate	FSC
	CDC	Systec	Sybase
	ADV/ORGA	Cisco	T-Mobile
	Interdata	3Com	Ixos
	Amdahl	Compaq	AMD
	Datev	Atlantic	Mobilcom
	Hewlett	Documentum	SGI
	Compagnie	Kombinat	Palm
	MDS	NEC	Oracles

	CW_{5kb}		
Date	1975	1988	2003
	Taylorix	Sony	AOL
	Memorex	Harris	Sony
	Packard	EDS	HPs
	Varian	SCO	T-Online
	Facit	ADR	SAS
	MBB	CDC	Vignette
	Inforex	Informix	3Com
	CTM	Norsk	Adobe
	Triumph-Adler	AEG	Navision
	Calcomp	DECs	Suns
	Olympia	VEB	Unisys
	AEG-Telefunken	Kodak	IT-Anbieter
	Datasaab	Vascom	Compaq
	Anker	Datev	Nokia
	Wang	Tandem	Infineon
	MS	Robotron	Acer
	AEG	Cullinet	Toshiba
	IBM-	MS	Corel
	ITT		BMW
			Documentum
			Gates-Company
			Intershop
			Lucent
			NEC
			SAPs
			Veritas
			Amazon
			Netscape
			Lexmark

Semantical Analysis CW_{5Kbu}, leading aggregated concepts within corpus segments and drill-down to term level:

Table 74: CW_{5kbu}, leading aggregated concepts within corpus segments and drill-down to term level

	CW5kbu		
Date	1975	1988	2003
Cc_CountThresU_Dim	IT	IT	IT
	Economy	Economy	Vendor
	Vendor	Vendor	Event
	Business	Business	Economy
	Geography	Science	Business
Cv_CountThresU_Dim	Event	Event	Currency
	OS	OS	Norm
	Vendor	Vendor	OS
	Name	Name	Profession
	Institute	Institute	ITProduct
	Business	Business	Vendor
	Geography	Geography	Customer
	Economy	Economy	ProgLanguage
	IT	IT	IT
	Currency	Currency	Economy
	Profession	Profession	Institute
	ITProduct	ITProduct	Geography
	Norm	Norm	Name
	Customer	Customer	Business
	Performance	Performance	Performance
	Science	Science	SocialFramework
	SocialFramework	SocialFramework	
	ProgLanguage	ProgLanguage	
Chipcom.TermFirstOcc	1987		
Chipcom.TermLastOcc	1998		
Cc_CountThresU_Vendor	IBM	IBM	IBM
		Microsoft	
Cv_CountThresU_Vendor	Telekom	Telekom	Novell
	3Com	Bertelsmann	Oracle
	Vodafone	Novell	SCO
	Infineon	SAS	Vodafone
	Oracle	Hyperion	Borland
	Ariba	Infineon	Fujitsu-Siemens
	Lycos	Toshiba	Microsofts
	T-Online	Mobilcom	Abb
	AMD	Siebel	Telekom
	Matsushita	Oracle	EDS
	Nokia	Pixelpark	Suse
	Microstrategy	Nokia	Siebel
	Worldcom	SAPs	Bea
	Honeywell	AOL	Cisco
	Univac	Freenet	Google
	Unidata	Bundespost	Symantec
	Kienzle	Amdahl	Ericsson
	Sperry	Unisys	Weblogic
	Singer	Wang	EMC
	Bundespost	Apollo	FSC
	Burroughs	Ashton-Tate	Sybase
	CDC	Systec	T-Mobile
	ADV/ORGA	Cisco	Ixos
	Interdata	3Com	AMD
	Amdahl	Compaq	Mobilcom
	Datev	Atlantic	SGI
	Hewlett	Documentum	Palm
	Compagnie	Kombinat	Oracles
	MDS	NEC	Sony
	Taylorix	Sony	AOL
	Memorex	EDS	Suns
	Packard	Harris	HPs
	Varian	SCO	T-Online
	Facit	ADR	Vignette
	MBB	CDC	Navision
	Inforex	DECs	SAS
	CTM	Informix	3Com
	Triumph-Adler	Norsk	Adobe
	Calcomp	AEG	IT-Anbieter
	Olympia	VEB	Unisys
	AEG-Telefunken	Kodak	Compaq
	Datasaab	Vascom	Nokia
	Anker	Datev	Infineon

	CW5kbu		
Date	1975	1988	2003
	Wang	Robotron	Acer
	MS	Tandem	Toshiba
	AEG	Cullinet	Corel
	IBM-	MS	BMW
	ITT		Documentum
	Centronics		Gates-Company
			Lucent
			Intershop
			NEC
			SAPs
			Veritas
			Amazon
			Netscape
			Lexmark
			MCI
			Hyperion

Semantical Analysis CW_{5Kbun}, leading aggregated concepts within corpus segments and drill-down to term level:

Table 75: CW_{5kbun}, leading aggregated concepts within corpus segments and drill-down to term level

	CW5kbun		
Date	1975	1988	2003
Cc_CountThresU_Dim	IT	IT	IT
Cv_CountThresU_Dim	Event	OS	Event
	Vendor	Vendor	ITProduct
	Institute	ProgLanguage	Currency
	Geography	SocialFramework	OS
	Business	Science	Vendor
	Customer	Business	Economy
	OS	ITProduct	Norm
	Name	Economy	IT
	IT	Event	Geography
	Economy	Profession	Business
	SocialFramework	Customer	Customer
	Profession	IT	Performance
		Name	Profession
		Geography	Name
		Performance	Science
Chipcom.TermFirstOcc	1994		
Chipcom.TermLastOcc	1995		
Cc_CountThresU_Vendor	-	-	-
Cv_CountThresU_Vendor	Telekom	Telekom	SCO
	3Com	Bertelsmann	HP
	Vodafone	Fujitsu	Novell
	Apple	Apple	Sun
	Ariba	Digital	Dell
	HP	DEC	Oracle
	Infineon	SAS	Vodafone
	Lycos	Dell	SAP
	Oracle	Hyperion	3Com
	T-Online	Infineon	Borland
	AMD	Novell	Fujitsu-Siemens
	Matsushita	SAP	Sharp
	Microstrategy	Sun	Suse
	Next	Toshiba	Ixos
	Nokia	IDV	SAG
	Worldcom	Mobilcom	Bea
	Kienzle	Amdahl	NCR
	Unidata	mbp	Hewlett-Packard
	CDC		Lexmark

Semantical Analysis CW_{5Kbun2}, leading aggregated concepts within corpus segments and drill-down to term level:

Table 76: CW_{5kbun2}, leading aggregated concepts within corpus segments and drill-down to term level

	CW_{5kbun2}		
Date	1975	1988	2003
Cc_CountThresU_Dim	IT	IT	IT
Cv_CountThresU_Dim	Event	Performance	Event
	Currency	ProgLanguage	Norm
	Vendor	Vendor	ITProduct
	Institute	Event	Vendor
	Geography	Currency	OS
	Business	OS	Currency
	Name	Science	Economy
	OS	ITProduct	IT
	SocialFramework	SocialFramework	Customer
	IT	Business	Institute
	Profession	Profession	Business
	Economy	Economy	Geography
	Customer	Name	Science
	Performance	Geography	ProgLanguage
		IT	Name
		Norm	Profession
			Performance
Chipcom.TermFirstOcc	1993		
Chipcom.TermLastOcc	1996		
Cc_CountThresU_Vendor	-	-	-
Cv_CountThresU_Vendor	Telekom	Telekom	Novell
	3Com	Bertelsmann	Sun
	Vodafone	Apple	Siemens
	Siemens	DEC	HP
	Apple	Digital	Borland
	Ariba	Fujitsu	Dell
	HP	Novell	Oracle
	Infineon	Dell	SAP
	Lycos	Hyperion	SCO
	Oracle	Infineon	Fujitsu-Siemens
	T-Online	SAS	Sharp
	AMD	SAP	Vodafone
	Matsushita	Sun	Gemini
	Microstrategy	Toshiba	Cisco
	Next	Siebel	Documentum
	Nokia	Oracle	Microsofts
	Worldcom	Mobilcom	EDS
	Inforex	Siemens	Digital
		Sony	
		Pixelpark	
		Stratus	
		SAPs	

Acknowledgements

This dissertation was initially inspired during the years 2000-2001 by many brainstorming discussions shared with Prof. Dr. Hans Gernert †[36] (Faculty of Business Informatics at the School of Business and Economics of the Humboldt University in Berlin). He widened my understanding of the knowledge domain “Business informatics” with critical questions related to its scientific development and such leading topics. After Prof. Gernert’s death Dr. Bernd Viehweger and Prof. Dr. Oliver Günther gave their support to enable the successful completion of my dissertational approach. Special help and inspiring discussions in corpus linguistical topics were provided by Prof. Dr. Anke Lüdeling and by Prof. Dr. Myra Spiliopoulou in the Data Mining research field.

[36] Prof. Hans Gernert unfortunately died after an accident in 2003.

Eidestattliche Erklärung

Ich bezeuge durch meine Unterschrift, dass meine Angaben über die bei der Abfassung meiner Dissertation benutzten Hilfsmittel, über die mir zuteil gewordene Hilfe sowie über frühere Begutachtungen meiner Dissertation in jeder Hinsicht der Wahrheit entsprechen.

Berlin, den 07.08.2007

Tobias Kalledat